Rapport d'étude

Modèle linéaire généralisé et choix de modèles

EMSBD6 - Bruno KUBECZKA

9 Juillet 2023

Abstract

Cette étude a pour objet la mise en pratique des techniques de régression logistique et de sélection de modèles dans le cadre de la prédiction du risque de pluie dans la ville de Bâle.

Contents

1 Introduction

1.1 Rapport et script R

Ce rapport présente la **démarche de l'étude** et l'**analyse des résultats obtenus**.

Il accompagne le script **projet_mlg.R** contenant l'ensemble du code ayant permis

- l'analyse exploratoire des jeux de données,
- l'identification des modèles
- la mesure des capacités prédictives
- la prédiction et la décision

L'organisation du code R et celle du rapport ont été alignées afin de faciliter les allers-retours entre code et analyse.

1.2 Approche selon la méthode Hold-Out

Nous avons à disposition 2 jeu de données :

- un jeu de données **meteo.train** de **1180 observations** pour lequel la variable d'intérêt ***pluie.demain*** est donnée
- un jeu de données **meteo.test** de **290 observations** pour lequel la variable d'intérêt ***pluie.demain*** n'est pas donnée

Considérons l'hypothèse que les données des 2 jeux de données sont issus d'un même jeu sur lequel un tirage aléatoire 75% / 25% a été réalisé. Etant donnés les index contenus dans les jeux de données fournis, l'hypothèse est raisonnable.

Nous pouvons alors aborder l'étude selon une **méthode de type Hold-Out**, à savoir:

- Le jeu de données **meteo.train** servira à une **phase d'entraînement et de validation**.

  Il contribuera à identifier le meilleur modèle dans l'optique de prédire de nouvelles valeurs.

  Le jeu de données **meteo.train** va être partitionné en **2 jeux de données distincts, tirés au hasard** selon un ratio 80/20.

  - Un **jeu de données d'entraînement des modèles** (Training Set - 80% des individus)
    Il permettra l'ajustement des paramètres des modèles candidats à la prédiction.
  - Un **jeu de données de validation des modèles** (Validation Set - 20% des individus)
    Il permettra de mesurer la capacité prédictive des modèles selon des critères vus au chapitre suivant.

    A l'issue de cette phase de validation, on conclura sur le modèle le plus à même de prédire les valeurs du jeu **meteo.test**

- Le jeu de données **meteo.test** (Holdout Set) servira à une **phase de test** du "meilleur" modèle entraîné et validé.

  Cette phase se limitera à la prédiction des valeurs binaires ***pluie.demain***, le résultat de test faisant l'objet de l'évaluation du projet.

1.3 Identification des modèles candidats à la prédiction

On débutera l'étude par une **analyse exploratoire**. Elle permettra une 1ère approche

- des corrélations entre les covariables et la variable d'intérêt ***pluie.demain***
- des corrélations entre les covariables elles-mêmes

L'analyse des covariables se fera par **famille de covariables** : température, humidité, pression nébulosité, vent.

A partir de l'analyse, il pourra être possible d'identifier une **sélection de covariables pertinentes** qui nous amènera à considérer **2 stratégies d'identification de modèles candidats** :

- une **1ère stratégie naïve** consistera
  - à ajuster un modèle complet incluant toutes les covariables fournies par le fichier **meteo.train** sans précaution aucune relative aux colinéarités identifiées entre les covariables
  - puis à optimiser ce modèle initial par des méthodes pas-à-pas (cf. ci-dessous)
- une **2ème stratégie basée sur l'analyse exploratoire** consistera
  - à ajuster un modèle complet exploitant les enseignements issus de l'étude préalable des covariables ; il s'agira de prendre en considération les colinéarités des covariables et d'introduire si besoin dans ce modèle des covariables transformées (amplitude, variables seuils booléennes).
  - puis à optimiser ce modèle initial par des méthodes pas-à-pas (cf. ci-dessous)

L'idée de mettre en compétition ces 2 stratégies est de comparer 2 approches, l'une basée sur l'utilisation des outils sans connaissance préalable des covariables, l'autre basée sur une connaissance plus fine des données.

Méthodes de sélection de modèle pas-à-pas

Pour chacune des stratégies, on envisagera les **méthodes pas-à-pas** suivantes

- **méthode descendante** depuis le modèle complet
- **méthode progressive descendante** depuis le modèle complet
- **méthode ascendante** depuis un modèle constant **pluie.demain** ~ **1** vers le modèle complet
- **méthode progressive ascendante** depuis un modèle constant **pluie.demain** ~ **1** vers le modèle complet

Chacune des méthodes sera basée sur la minimisation du score AIC.

IMPORTANT : reproductibilité des résultats

Dans un souci de reproductibilité des résultats présentés dans ce rapport, le vecteur aléatoire permettant de séparer les données d'entraînement et de validation au sein du jeu **meteo.train** est sauvegardé dans un fichier **ref.training_validation.rdata** (disponible sur le dépôt du projet) :

- si le fichier **ref.training_validation.rdata** est présent dans le répertoire du script R, alors le fichier est chargé et utilisé
- si le fichier est absent, un nouveau vecteur aléatoire training-validation est généré, respectant le ratio 80% / 20%.

1.4 Critères de sélection

Chaque modèle candidat sera indexé dans une table de résultats (RESULTS) et pour chacun on relèvera :

- ses covariables
- son **score AIC** issu de l'ajustement des paramètres via la méthode **glm**
- son **auc** (Area Under the Curve) et sa **précision** issus de la prédiction du jeu de données de validation :
  - L'AUC est issu de l'utilisation de la fonction ROCR::performance
  - La précision est calculée selon la formule $precision = \frac{True\ Positive + True\ Negative}{n}$
- sa **perte** issue de la prédiction du jeu de données de validation :
  La perte est calculée selon la formule $perte = \frac{1}{n_{val}} \sum_{i=1}^{n_{val}} |\tilde{p}_i - Y_i|$
  où $\tilde{p}_i$ est la probabilité $P(Y_i = 1)$ prédite et Y_i est la valeur observée
- ses **déviances** par rapport aux modèles nuls et saturés (p-valeurs issues des tests de chi2)

A noter, dans le cadre de cette étude

- on retient le **score AIC** comme méthode de sélection de variables.
  La finalité étant la prédiction plutôt que l'inférence, l'optimisation du nombre de covariables n'est pas un sujet du moment que le modèle prédit de façon performante. Il n'est donc pas nécessaire ici de considérer l'avantage du score **BIC** de pénaliser les modèles au nombre de covariables élevé.
- Le **Cp de Mallows** aurait pu apporter une approche pertinente basée sur la minimisation des résidus. Néanmoins, les fonctions disponibles sous R, en l'occurrence leaps::regsubsets ne sont pas adaptées à la régression logistique.
  La prise en compte de la **perte** pour les différents modèles permettra d'apporter un éclairage similaire.
- La décision des valeurs booléennes **pluie.demain** suite à la prédiction des probabilités P(Yi=1) se fera selon un **seuil** qui sera optimisé pour chaque modèle sur le jeu d'entraînement. Les jeux de validation et de test étant issus du même jeu de données initial, on considérera que cette optimisation s'applique à la totalité des données.

2 Préparation des données

Une fois les données **meteo.train** et **meteo.test** chargées, les opérations suivantes sont effectuées :

- Toutes les variables sont renommées pour plus de lisibilité.

  Ci-dessous le tableau de correspondance des noms de colonnes.

- les colonnes ***Year***, ***Day***, ***Hour*** et ***Minute*** ont été retirées.

- la donnée ***Month*** a été convertie en **variable catégorielle**

- Des variables **amplitude** (max-min) sont ajoutées en fin de tableau pour toutes les variables sous la forme **mean/min/max**.

- Des **variables booléennes** sont ajoutées en fin de tableau sous la forme pour les données ***precipitation, snowfall et sunshine***.

  Elles répondent aux questions

  - "a-t-il plu aujourd'hui ?" (***precipitation_bool***)
  - "a-t-il neigé aujourd'hui ?" (***snowfall_bool***)
  - "y-a-t-il eu du soleil aujourd'hui ?" (***sunshine_bool***)

Table 1: Tableau de correspondances

original	new
Month	month
Temperature.daily.mean..2.m.above.gnd.	temperature.mean
Relative.Humidity.daily.mean..2.m.above.gnd.	humidity.mean
Mean.Sea.Level.Pressure.daily.mean..MSL.	pressure.mean
Total.Precipitation.daily.sum..sfc.	precipitation
Snowfall.amount.raw.daily.sum..sfc.	snowfall
Total.Cloud.Cover.daily.mean..sfc.	total.cloud.mean
High.Cloud.Cover.daily.mean..high.cld.lay.	high.cloud.mean
Medium.Cloud.Cover.daily.mean..mid.cld.lay.	med.cloud.mean
Low.Cloud.Cover.daily.mean..low.cld.lay.	low.cloud.mean
Sunshine.Duration.daily.sum..sfc.	sunshine
Shortwave.Radiation.daily.sum..sfc.	radiation
Wind.Speed.daily.mean..10.m.above.gnd.	wind.speed.mean.10
Wind.Direction.daily.mean..10.m.above.gnd.	wind.dir.10
Wind.Speed.daily.mean..80.m.above.gnd.	wind.speed.mean.80
Wind.Direction.daily.mean..80.m.above.gnd.	wind.dir.80
Wind.Speed.daily.mean..900.mb.	wind.speed.mean.900
Wind.Direction.daily.mean..900.mb.	wind.dir.900
Wind.Gust.daily.mean..sfc.	wind.gust.mean
Temperature.daily.max..2.m.above.gnd.	temperature.max
Temperature.daily.min..2.m.above.gnd.	temperature.min
Relative.Humidity.daily.max..2.m.above.gnd.	humidity.max
Relative.Humidity.daily.min..2.m.above.gnd.	humidity.min
Mean.Sea.Level.Pressure.daily.max..MSL.	pressure.max
Mean.Sea.Level.Pressure.daily.min..MSL.	pressure.min
Total.Cloud.Cover.daily.max..sfc.	total.cloud.max
Total.Cloud.Cover.daily.min..sfc.	total.cloud.min
High.Cloud.Cover.daily.max..high.cld.lay.	high.cloud.max
High.Cloud.Cover.daily.min..high.cld.lay.	high.cloud.min
Medium.Cloud.Cover.daily.max..mid.cld.lay.	med.cloud.max
Medium.Cloud.Cover.daily.min..mid.cld.lay.	med.cloud.min
Low.Cloud.Cover.daily.max..low.cld.lay.	low.cloud.max
Low.Cloud.Cover.daily.min..low.cld.lay.	low.cloud.min
Wind.Speed.daily.max..10.m.above.gnd.	wind.speed.max.10
Wind.Speed.daily.min..10.m.above.gnd.	wind.speed.min.10
Wind.Speed.daily.max..80.m.above.gnd.	wind.speed.max.80
Wind.Speed.daily.min..80.m.above.gnd.	wind.speed.min.80
Wind.Speed.daily.max..900.mb.	wind.speed.max.900
Wind.Speed.daily.min..900.mb.	wind.speed.min.900
Wind.Gust.daily.max..sfc.	wind.gust.max
Wind.Gust.daily.min..sfc.	wind.gust.min

3 Analyse exploratoire

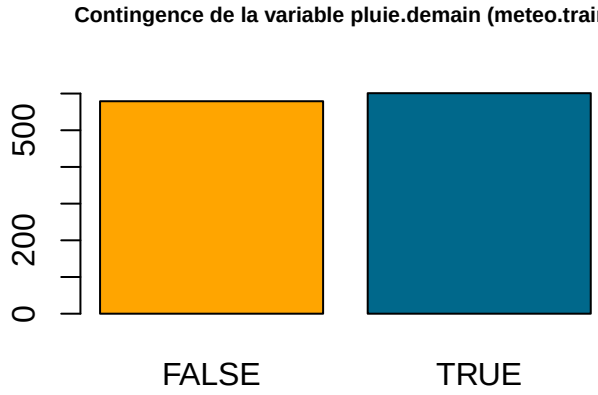
L'analyse exploratoire a pour objet

- de considérer l'impact de la variable ***pluie.demain*** sur la distribution des covariablers
- de mesurer les corrélations au sein d'une même famille ***mean/min/max/amplitude***
- de considérer les corrélations entre familles de covariables

3.1 Variable d'intérêt pluie.demain

Contingence de "pluie.demain" sur le jeu de données meteo.train

```
##
## FALSE  TRUE
##   579   601
```



On note que les valeurs ***pluie.demain*** sont équitablement réparties entre les valeurs TRUE / FALSE.

Aucun effet de sur-représentation d'une des valeurs n'est à considérer.

3.2 Analyse des covariables mean/min/max

Pour chaque donnée composée de valeurs moyenne / minimale / maximale

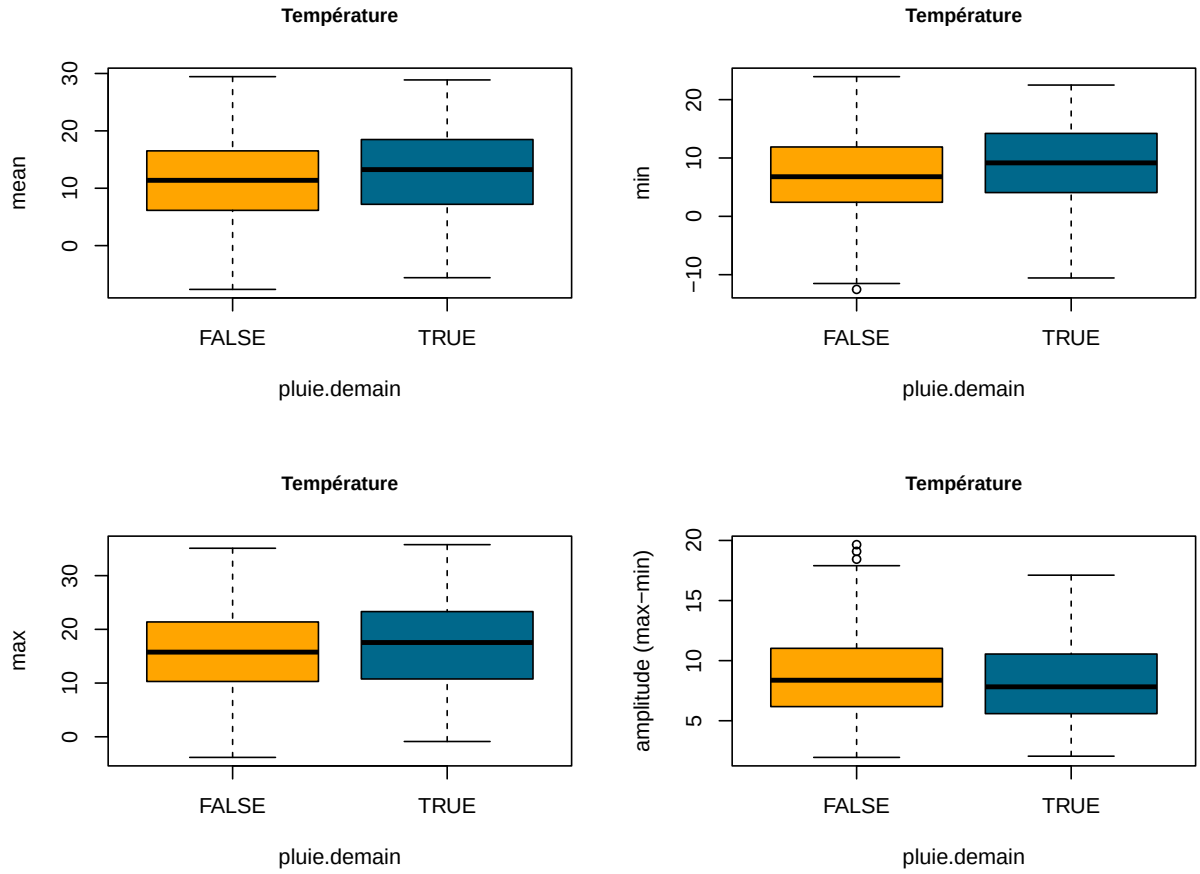
- on ajoute une variable **amplitude**
- on visualise le schéma de corrélation entre les 4 covariables
- on détermine les covariables les plus à même de représenter la famille dans le cadre d'un modèle où on aurait sélectionné les covariables (stratégie 2).

3.2.1 Température

covariables considérées

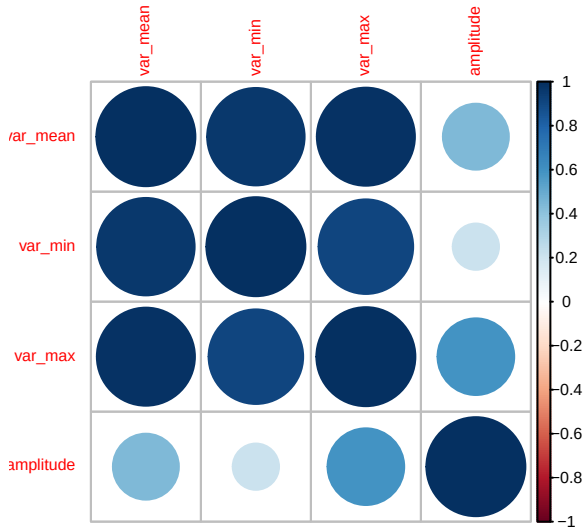
```
var_main="Température"
var_mean=dat.meteo.train$temperature.mean
var_min=dat.meteo.train$temperature.min
var_max=dat.meteo.train$temperature.max
```

Distribution des covariables selon pluie.demain



Corrélations entre covariables

```
##               var_mean   var_min   var_max amplitude
## var_mean   1.0000000 0.9683863 0.9806656 0.4432349
## var_min    0.9683863 1.0000000 0.9121400 0.2183299
## var_max    0.9806656 0.9121400 1.0000000 0.5991378
## amplitude 0.4432349 0.2183299 0.5991378 1.0000000
```



Analyse

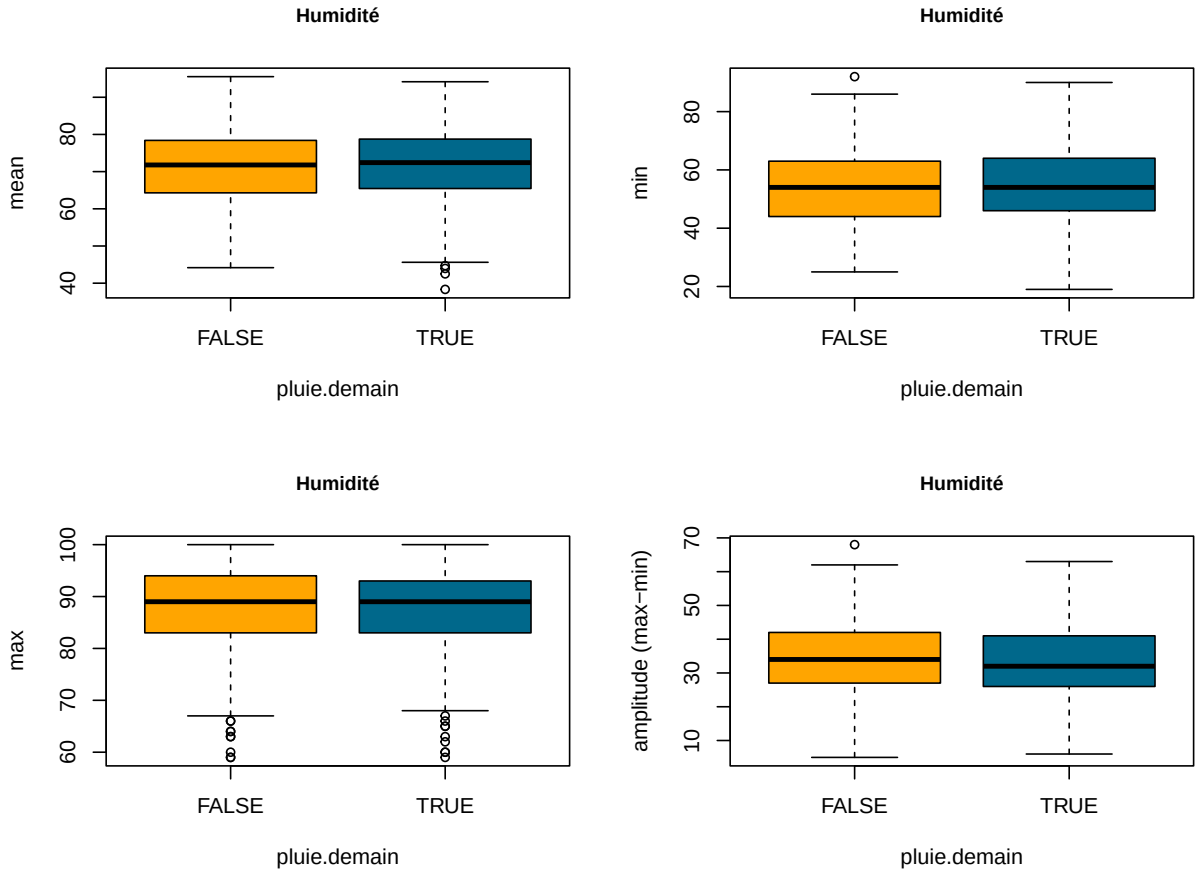
- L'impact de ***pluie.demain*** sur la distribution des covariables est **réel mais peu marqué**
  - les valeurs ***mean/min/max*** plus grandes **augmentent** le risque de pluie
  - les valeurs d'***amplitude*** plus petites **augmentent** le risque de pluie
- Corrélations
  - ***mean/min/max*** sont **fortement corrélées** entre elles
  - l'***amplitude*** est moyennement corrélée avec ***mean/max***
  - l'***amplitude*** est faiblement corrélée avec ***min***
- Dans le cadre de la sélection de variables (stratégie 2), on retiendra le produit **temperature.amplitude x temperature.min**

3.2.2 Humidité relative

covariables considérées

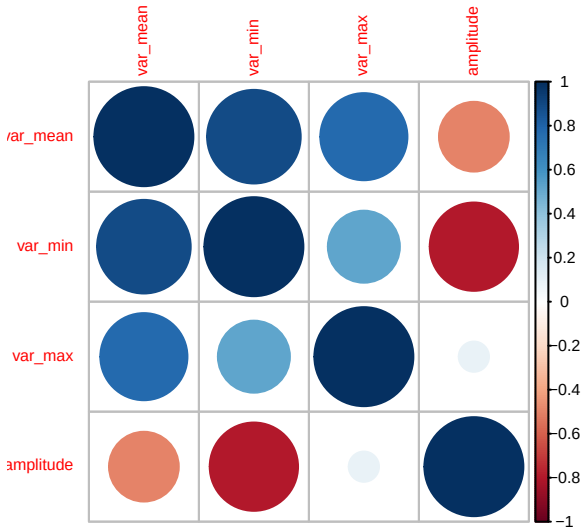
```
var_main="Humidité"
var_mean=dat.meteo.train$humidity.mean
var_min=dat.meteo.train$humidity.min
var_max=dat.meteo.train$humidity.max
```

Distribution des covariables selon pluie.demain



Corrélations entre covariables

```
##               var_mean    var_min    var_max   amplitude
## var_mean    1.0000000  0.8939695 0.77390562 -0.49621785
## var_min     0.8939695  1.0000000 0.52914005 -0.79544270
## var_max     0.7739056  0.5291400 1.00000000  0.09333577
## amplitude -0.4962178 -0.7954427 0.09333577  1.00000000
```



Analyse

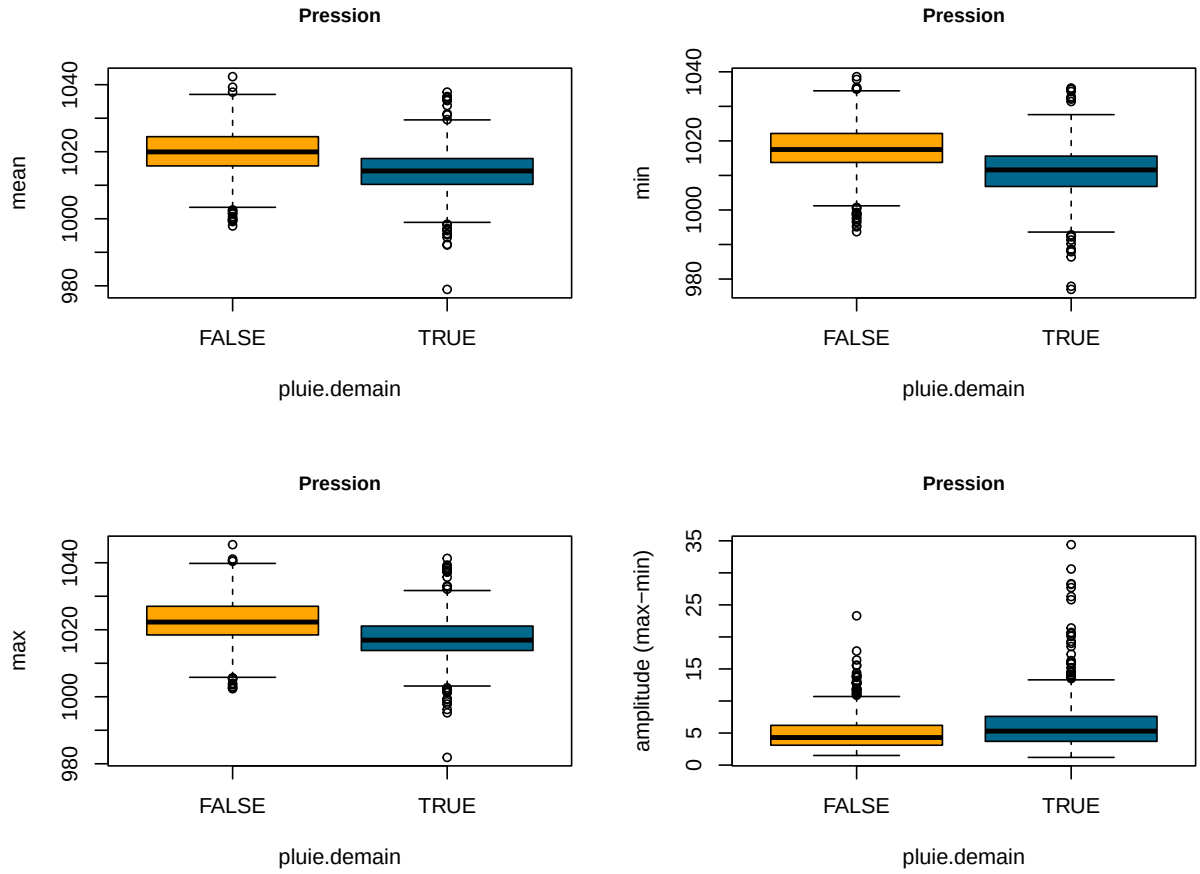
- S'il y a impact de ***pluie.demain*** sur la distribution des covariables, il est **très mesuré**
- Corrélations
  - ***mean/min/max*** sont **fortement corrélées** entre elles
  - l'***amplitude*** est moyennement corrélée avec ***mean/min***
  - l'***amplitude*** est faiblement corrélée avec ***max***
- Dans le cadre de la sélection de variables (stratégie 2), on retiendra le produit **humidity.amplitude x humidity.max**

3.2.3 Pression

covariables considérées

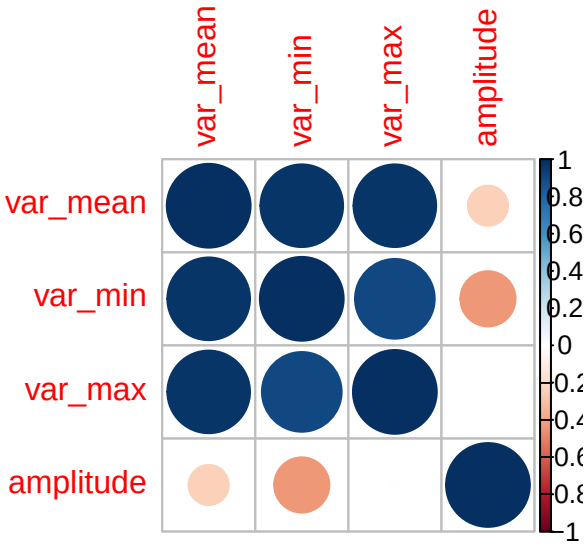
```
var_main="Pression"
var_mean=dat.meteo.train$pressure.mean
var_min=dat.meteo.train$pressure.min
var_max=dat.meteo.train$pressure.max
```

Distribution des covariables selon pluie.demain



Corrélations entre covariables

```
##              var_mean    var_min      var_max    amplitude
## var_mean    1.0000000  0.9735478  0.972200761 -0.230220060
## var_min     0.9735478  1.0000000  0.904746052 -0.434919125
## var_max     0.9722008  0.9047461  1.000000000 -0.009935123
## amplitude -0.2302201 -0.4349191 -0.009935123  1.000000000
```



Analyse

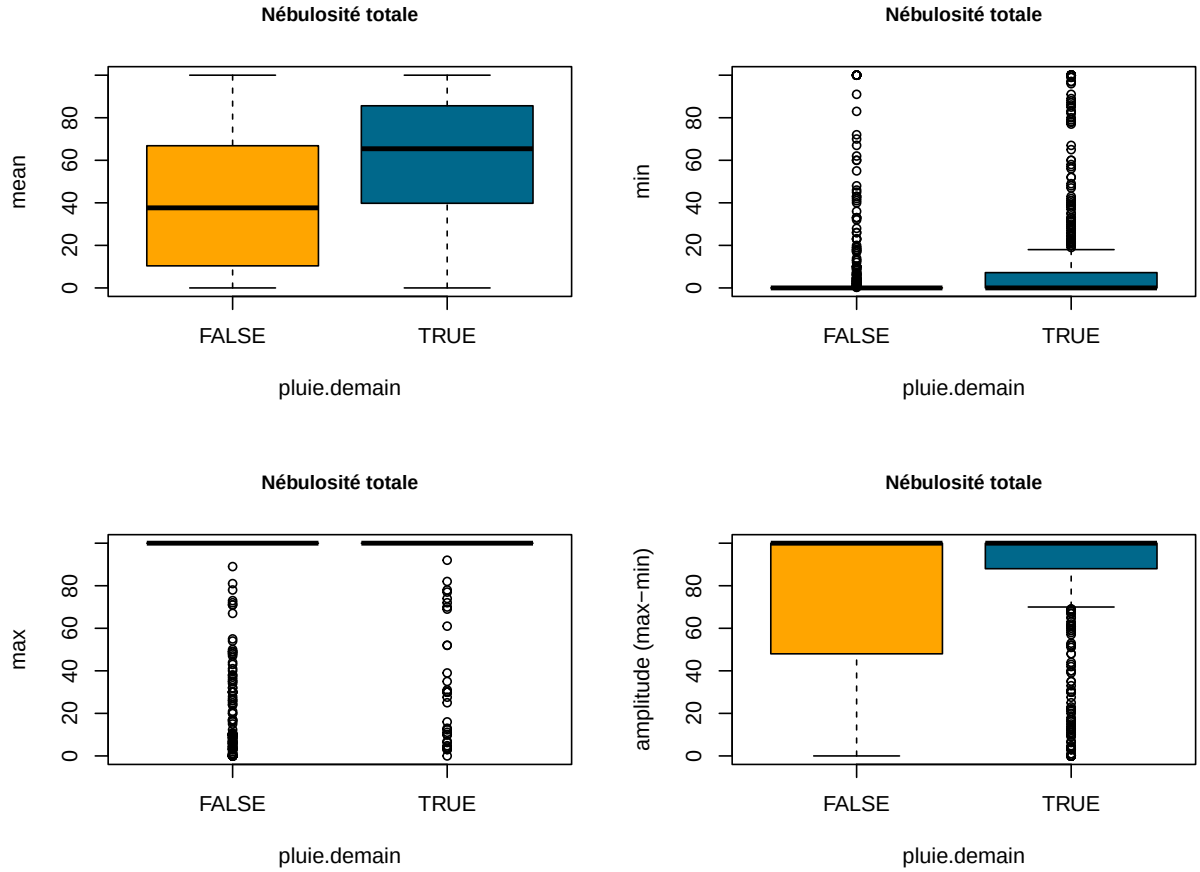
- L'impact de ***pluie.demain*** sur la distribution des covariables est **réel et relativement bien marqué**
  - les valeurs ***mean/min/max*** plus grandes **diminuent** le risque de pluie
  - les valeurs d'***amplitude*** plus grandes **augmentent** le risque de pluie
- Corrélations
  - ***mean/min/max*** sont **fortement corrélées** entre elles
  - l'***amplitude*** est moyennement corrélée avec ***mean/min***
  - l'***amplitude*** n'est pas corrélée à ***max***
- Dans le cadre de la sélection de variables (stratégie 2), on retiendra le produit **pressure.amplitude x pressure.max**

3.2.4 Nébulosité totale

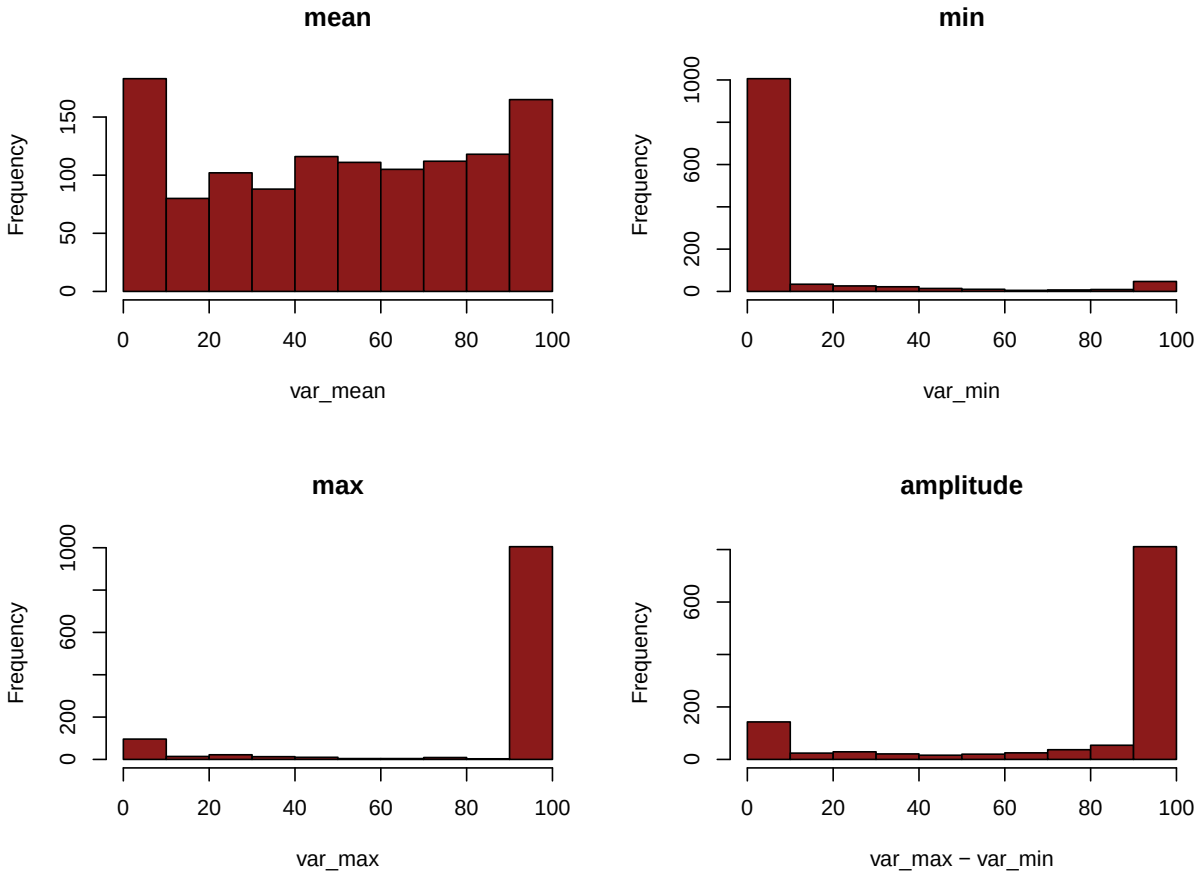
covariables considérées

```
var_main="Nébulosité totale"
var_mean=dat.meteo.train$total.cloud.mean
var_min=dat.meteo.train$total.cloud.min
var_max=dat.meteo.train$total.cloud.max
```

Distribution des covariables selon pluie.demain

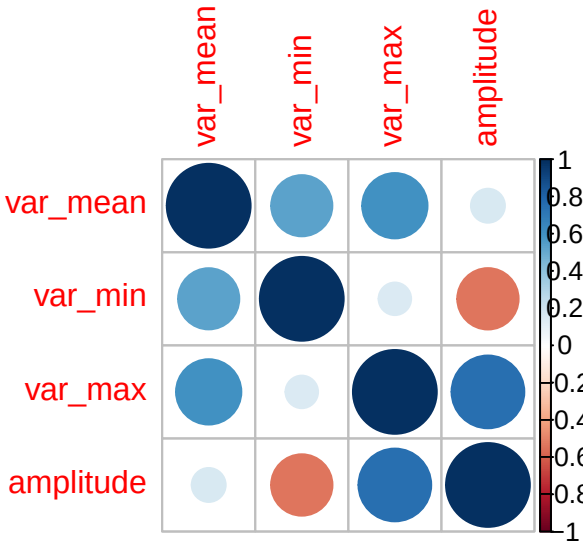


Histogramme des covariables



Corrélations entre covariables

```
##              var_mean    var_min   var_max  amplitude
## var_mean   1.0000000  0.5329495 0.6074140  0.1642736
## var_min    0.5329495  1.0000000 0.1509833 -0.5360951
## var_max    0.6074140  0.1509833 1.0000000  0.7535390
## amplitude 0.1642736 -0.5360951 0.7535390  1.0000000
```



Analyse

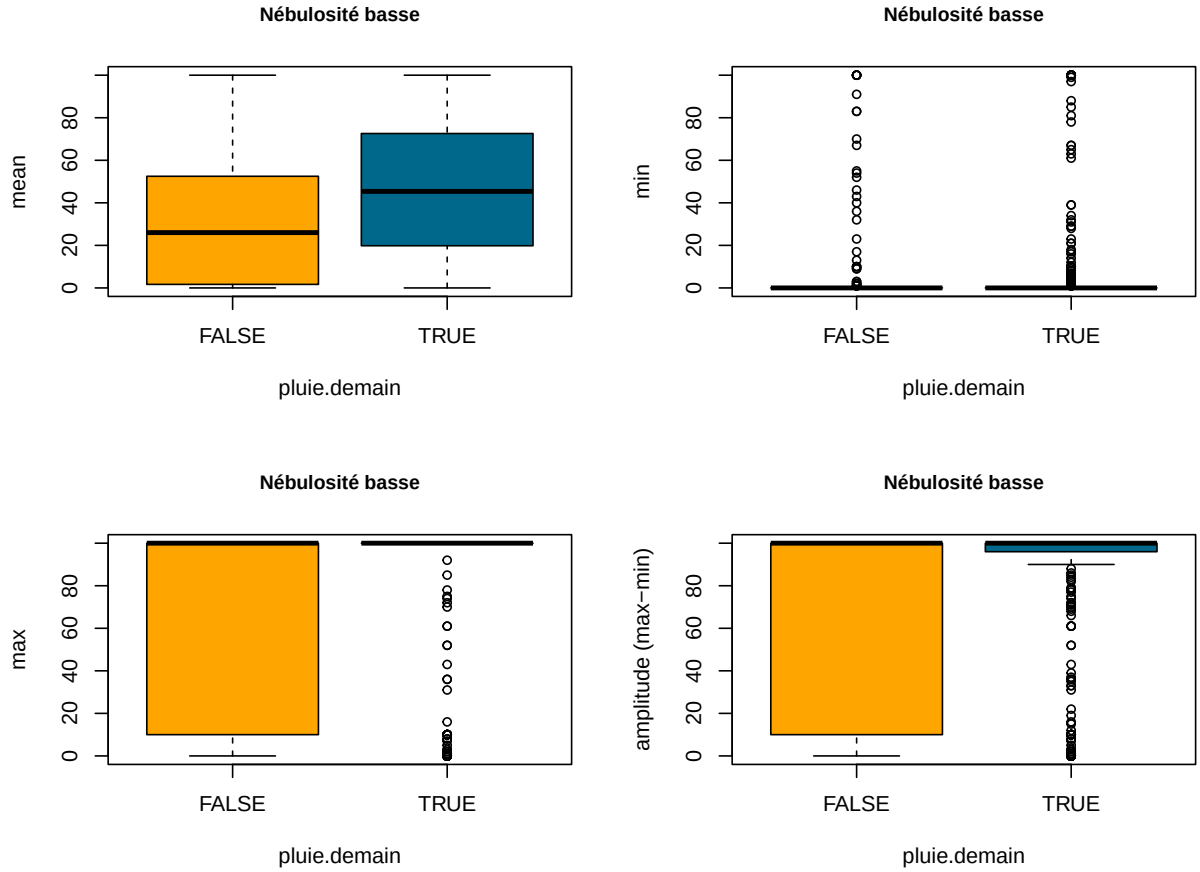
- L'impact de ***pluie.demain*** sur la distribution des **min/max** n'est pas déterminant
  Néanmoins, les valeurs ***mean/amplitude*** plus grandes **augmentent** le risque de pluie
- Corrélations
  - ***min/max*** sont **plutôt positivement corrélées** avec **mean**
  - ***min*** et ***max*** sont faiblement corrélées entre elles
  - l'***amplitude*** est fortement corrélée avec ***min/max***
  - l'***amplitude*** est faiblement corrélée avec ***mean***
- Dans le cadre de la sélection de variables (stratégie 2), on retiendra le produit **total.cloud.amplitude x total.cloud.mean**

3.2.5 Nébulosité basse

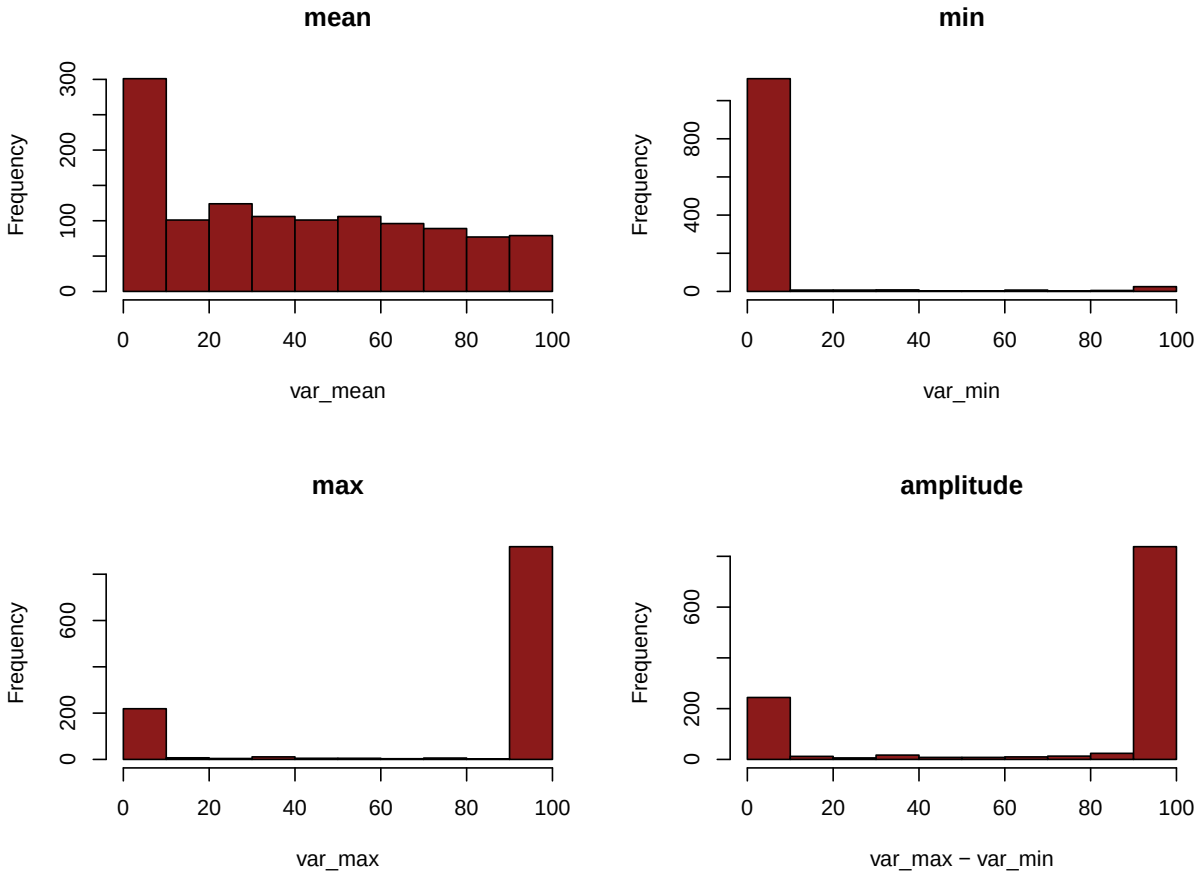
covariables considérées

```
var_main="Nébulosité basse"
var_mean=dat.meteo.train$low.cloud.mean
var_min=dat.meteo.train$low.cloud.min
var_max=dat.meteo.train$low.cloud.max
```

Distribution des covariables selon pluie.demain

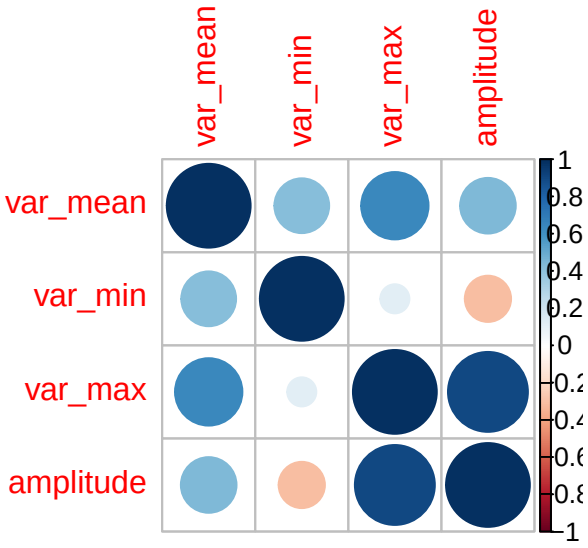


Histogramme de covariables



Corrélations entre covariables

```
##              var_mean    var_min   var_max  amplitude
## var_mean   1.0000000  0.4273706 0.6460351  0.4405601
## var_min    0.4273706  1.0000000 0.1203414 -0.3042092
## var_max    0.6460351  0.1203414 1.0000000  0.9090733
## amplitude 0.4405601 -0.3042092 0.9090733  1.0000000
```



Analyse

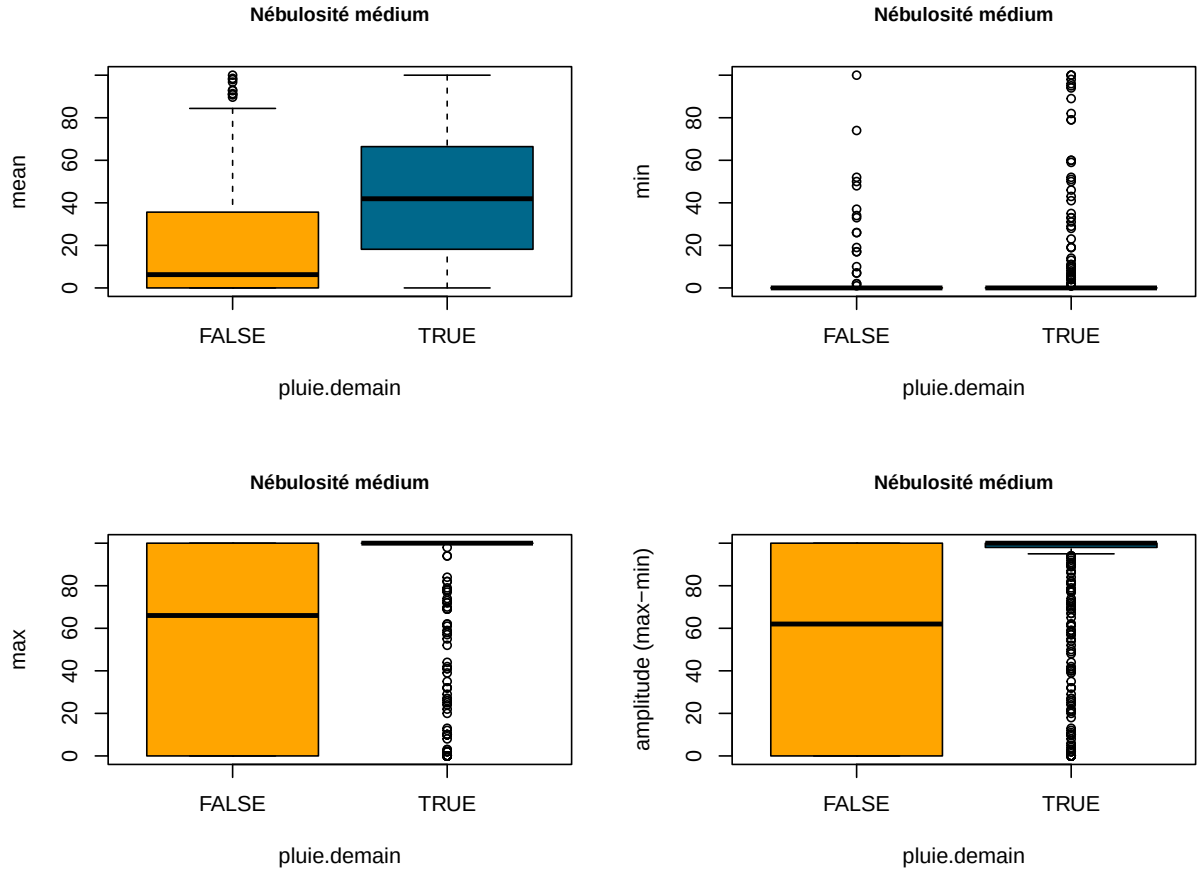
- L'impact de ***pluie.demain*** sur la distribution des **min** n'est pas déterminant
  Néanmoins, les valeurs ***mean/amplitude/max*** plus grandes **augmentent** le risque de pluie
- Corrélations
  - ***min/max*** sont **plutôt positivement corrélées** avec **mean**
  - ***min*** et ***max*** sont faiblement corrélées entre elles
  - l'***amplitude*** est fortement corrélée avec ***mean/max***
  - l'***amplitude*** est faiblement corrélée avec ***mean***
- Dans le cadre de la sélection de variables (stratégie 2), on retiendra le produit **low.cloud.amplitude x low.cloud.min**

3.2.6 Nébulosité medium

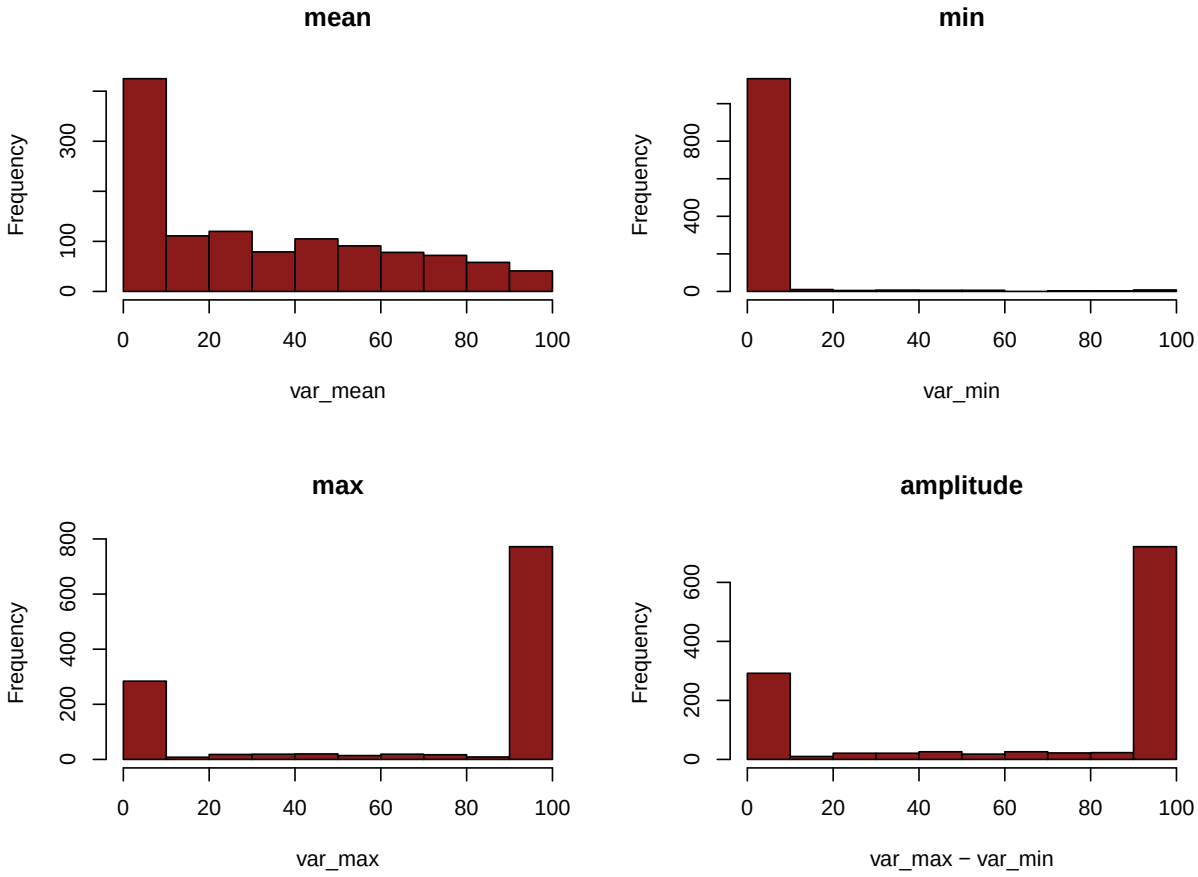
covariables considérées

```
var_main="Nébulosité médium"
var_mean=dat.meteo.train$med.cloud.mean
var_min=dat.meteo.train$med.cloud.min
var_max=dat.meteo.train$med.cloud.max
```

Distribution des covariables selon pluie.demain

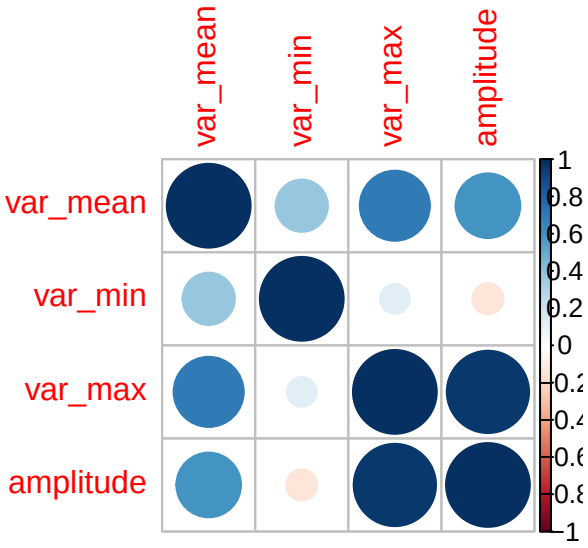


Histogramme des covariables



Corrélations entre covariables

```
##              var_mean    var_min   var_max  amplitude
## var_mean   1.0000000  0.3896551 0.7000574  0.5957026
## var_min    0.3896551  1.0000000 0.1265068 -0.1385866
## var_max    0.7000574  0.1265068 1.0000000  0.9648614
## amplitude 0.5957026 -0.1385866 0.9648614  1.0000000
```



Analyse

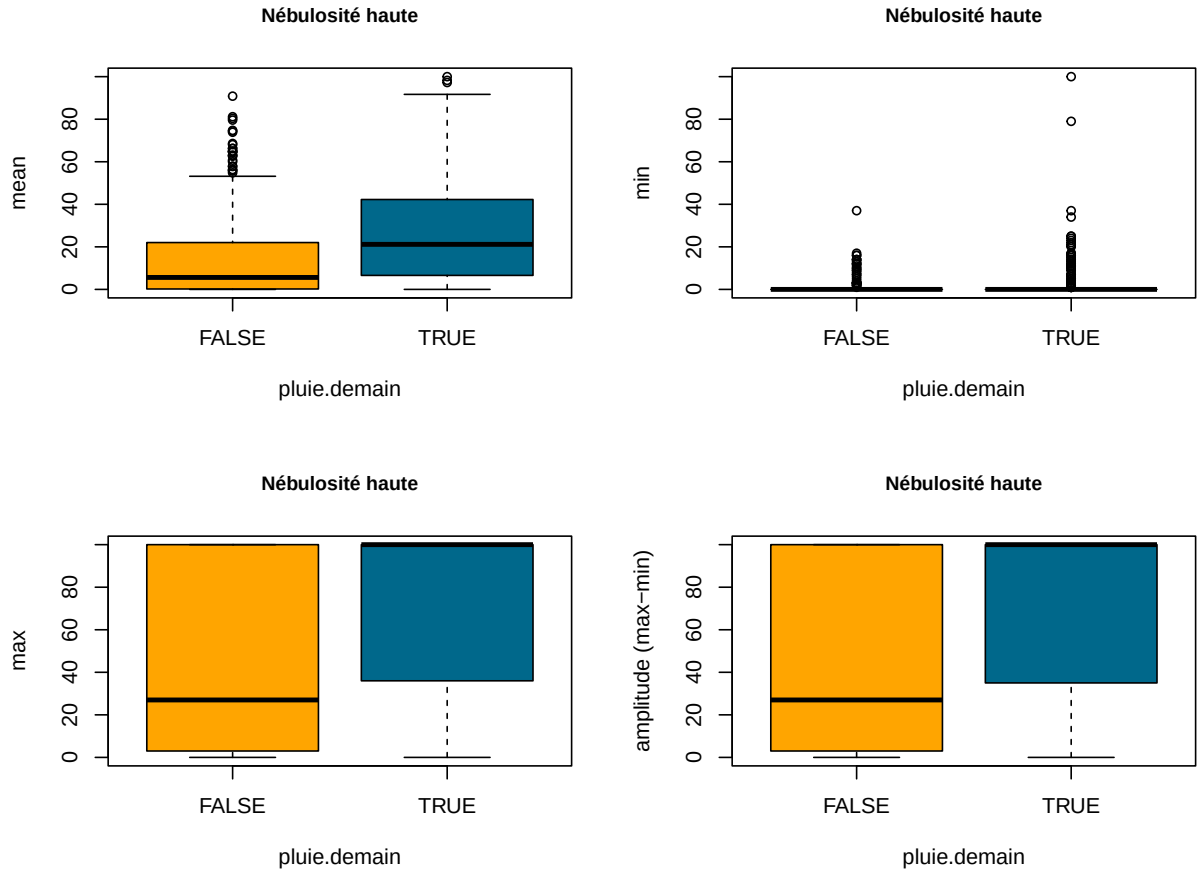
- L'impact de ***pluie.demain*** sur la distribution des **min** n'est pas déterminant
  Néanmoins, les valeurs ***mean/amplitude/max*** plus grandes **augmentent** le risque de pluie
- Corrélations
  - ***min/max*** sont **plutôt positivement corrélées** avec **mean**
  - ***min*** et ***max*** sont faiblement corrélées entre elles
  - l'***amplitude*** est fortement corrélée avec ***mean/max***
  - l'***amplitude*** est faiblement corrélée avec ***min***
- Dans le cadre de la sélection de variables (stratégie 2), on retiendra le produit **med.cloud.amplitude x med.cloud.min**

3.2.7 Nébulosité haute

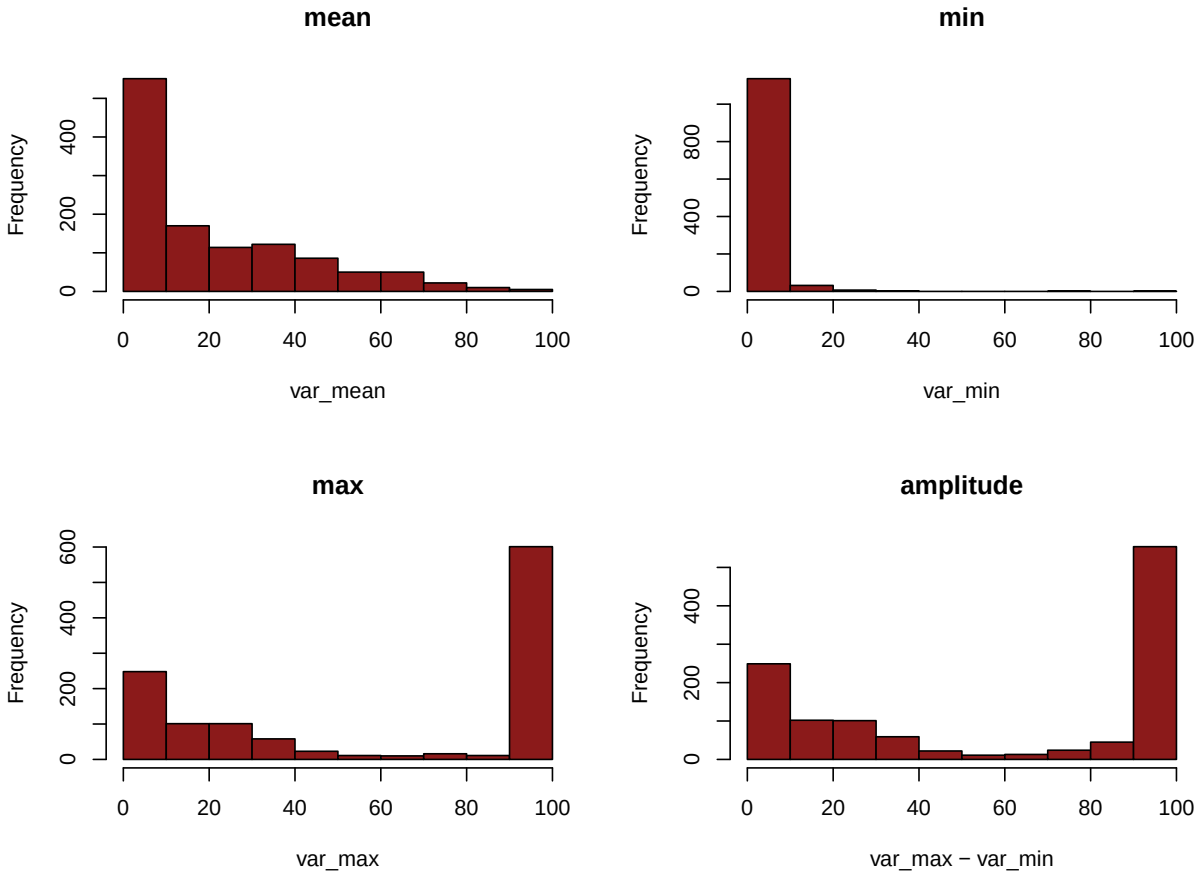
covariables considérées

```
var_main="Nébulosité haute"
var_mean=dat.meteo.train$high.cloud.mean
var_min=dat.meteo.train$high.cloud.min
var_max=dat.meteo.train$high.cloud.max
```

Distribution des covariables selon pluie.demain

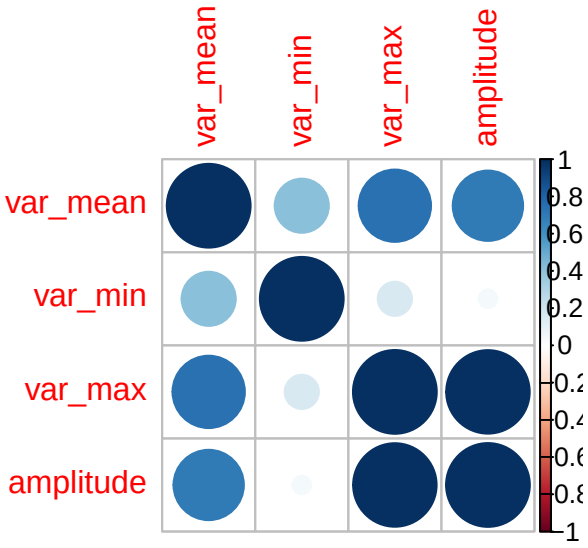


Histogramme des covariables



Corrélations entre covariables

```
##            var_mean    var_min   var_max  amplitude
## var_mean  1.0000000 0.41745207 0.7453130 0.70505473
## var_min   0.4174521 1.00000000 0.1665816 0.04919213
## var_max   0.7453130 0.16658163 1.0000000 0.99302842
## amplitude 0.7050547 0.04919213 0.9930284 1.00000000
```



Analyse

- L'impact de ***pluie.demain*** sur la distribution des **min** n'est pas déterminant
  Néanmoins, les valeurs ***mean/amplitude/max*** plus grandes **augmentent** le risque de pluie
- Corrélations
  - ***min/max*** sont **plutôt positivement corrélées** avec **mean**
  - ***min*** et ***max*** sont faiblement corrélées entre elles
  - l'***amplitude*** est fortement corrélée avec ***mean/max***
  - l'***amplitude*** est faiblement corrélée avec ***min***
- Dans le cadre de la sélection de variables (stratégie 2), on retiendra le produit **high.cloud.amplitude x high.cloud.min**

3.2.8 Corrélation entre nébulosités

```
##                  low.cloud.mean med.cloud.mean high.cloud.mean
## low.cloud.mean       1.0000000      0.5737378       0.2364875
## med.cloud.mean       0.5737378      1.0000000       0.6981242
## high.cloud.mean      0.2364875      0.6981242       1.0000000
```

```
##                 low.cloud.min med.cloud.min high.cloud.min
## low.cloud.min       1.0000000     0.3522855      0.1127867
## med.cloud.min       0.3522855     1.0000000      0.3636318
## high.cloud.min      0.1127867     0.3636318      1.0000000
```

```
##                 low.cloud.max med.cloud.max high.cloud.max
## low.cloud.max       1.0000000     0.3759214      0.1913910
## med.cloud.max       0.3759214     1.0000000      0.6654278
## high.cloud.max      0.1913910     0.6654278      1.0000000
```

```
##                       low.cloud.amplitude med.cloud.amplitude
## low.cloud.amplitude             1.0000000           0.3277003
## med.cloud.amplitude             0.3277003           1.0000000
## high.cloud.amplitude            0.1600297           0.6364111
##                       high.cloud.amplitude
## low.cloud.amplitude              0.1600297
## med.cloud.amplitude              0.6364111
## high.cloud.amplitude             1.0000000
```

Analyse

Les corrélations entre les différentes nébulosités sont raisonnables.

On prendra le partie de toutes les inclure toutes dans le cadre de la sélection de variables (stratégie 2).

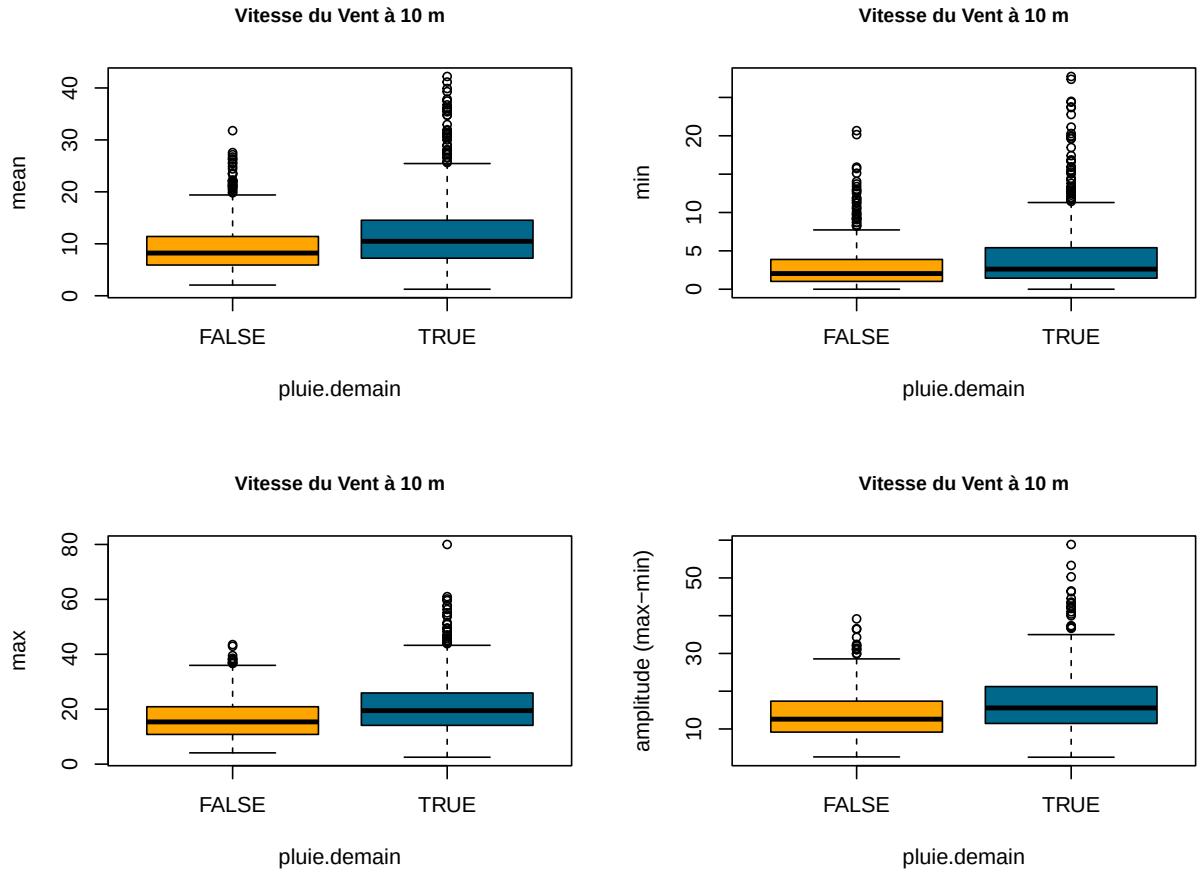
3.2.9 Vitesse et sens du vent à 10 m

3.2.9.1 Vitesse du vent :

covariables considérées

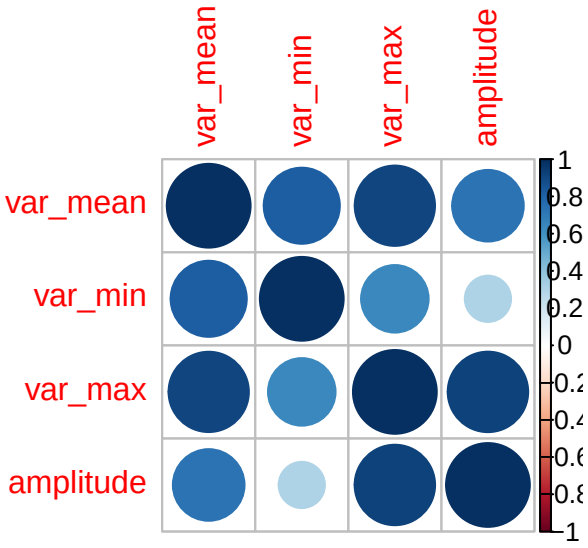
```
var_main="Vitesse du Vent à 10 m"
var_mean=dat.meteo.train$wind.speed.mean.10
var_min=dat.meteo.train$wind.speed.min.10
var_max=dat.meteo.train$wind.speed.max.10
```

Distribution des covariables selon pluie.demain



Corrélations entre covariables

```
##               var_mean   var_min   var_max amplitude
## var_mean   1.0000000 0.8250362 0.9185578 0.7308883
## var_min    0.8250362 1.0000000 0.6498235 0.3051279
## var_max    0.9185578 0.6498235 1.0000000 0.9221170
## amplitude 0.7308883 0.3051279 0.9221170 1.0000000
```



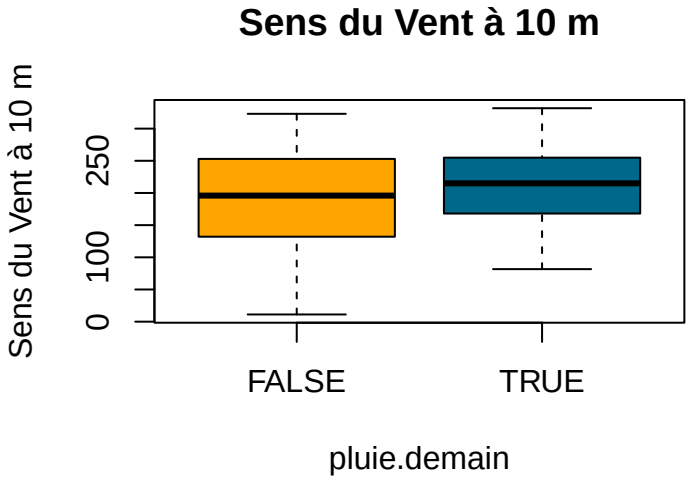
Analyse

- L'impact de ***pluie.demain*** sur la distribution des covariables est **réel mais peu marqué**
  - les valeurs ***mean/min/max*** plus grandes **augmentent** le risque de pluie
  - les valeurs d'***amplitude*** plus grandes **augmentent** le risque de pluie
- Corrélations
  - ***mean/min/max*** sont **fortement corrélées** entre elles
  - l'***amplitude*** est fortement corrélée avec ***mean/max***
  - l'***amplitude*** est faiblement corrélée avec ***min***
- Dans le cadre de la sélection de variables (stratégie 2), on retiendra le produit **wind.speed.amplitude.10 x wind.speed.min.10**

3.2.9.2 Sens du vent :

covariable considérée

Distribution selon pluie.demain



Analyse

- L'impact de ***pluie.demain*** sur la distribution est **réel**
- Dans le cadre de la sélection de variables (stratégie 2), on retiendra le produit **wind.speed.amplitude.10 x wind.speed.min.10 x wind.dir.10**

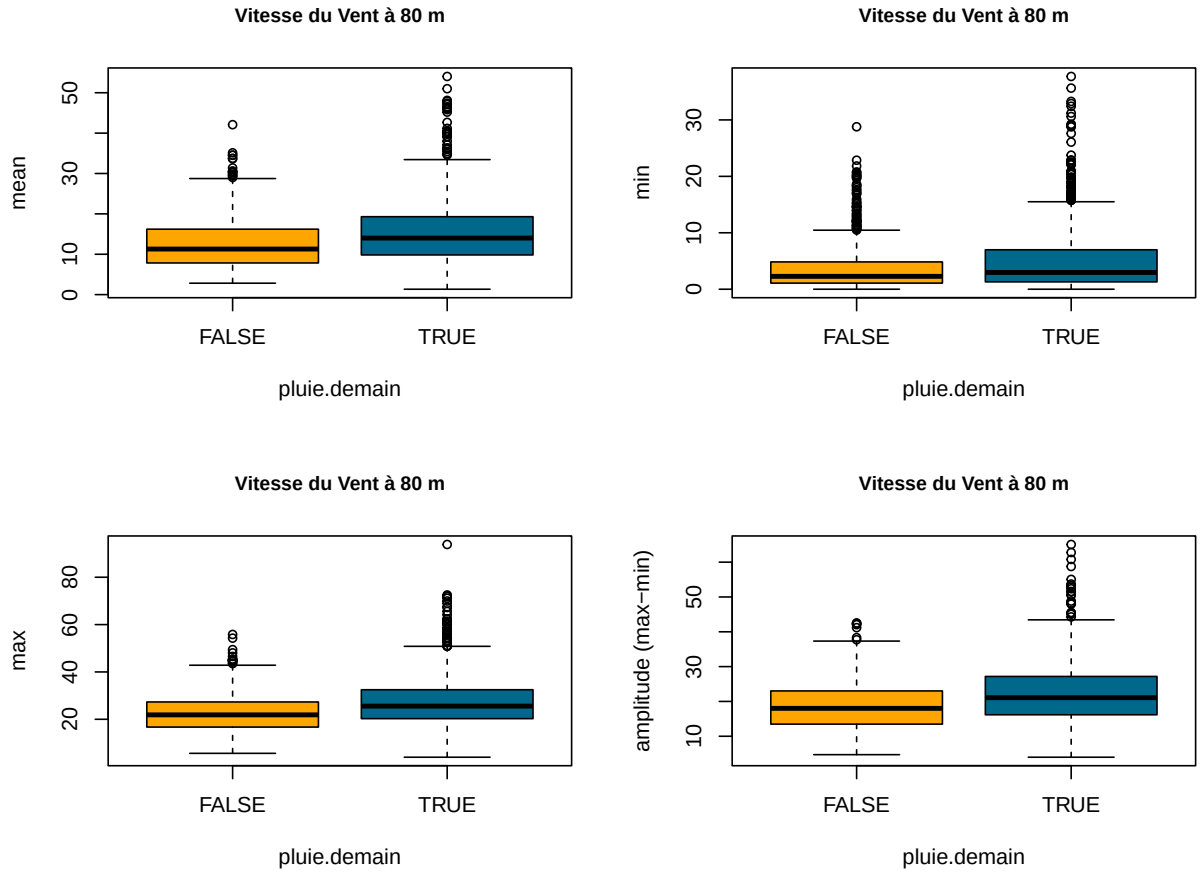
3.2.10 Vitesse et sens du vent à 80 m

3.2.10.1 Vitesse du vent :

covariables considérées

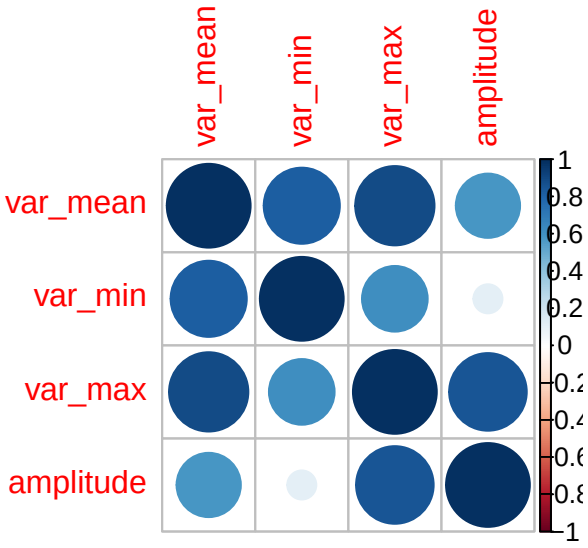
```
var_main="Vitesse du Vent à 80 m"
var_mean=dat.meteo.train$wind.speed.mean.80
var_min=dat.meteo.train$wind.speed.min.80
var_max=dat.meteo.train$wind.speed.max.80
```

Distribution des covariables selon pluie.demain



Corrélations entre covariables

```
##               var_mean   var_min   var_max amplitude
## var_mean   1.0000000 0.8251264 0.8953142 0.5879983
## var_min    0.8251264 1.0000000 0.6137479 0.1199484
## var_max    0.8953142 0.6137479 1.0000000 0.8574200
## amplitude 0.5879983 0.1199484 0.8574200 1.0000000
```



Analyse

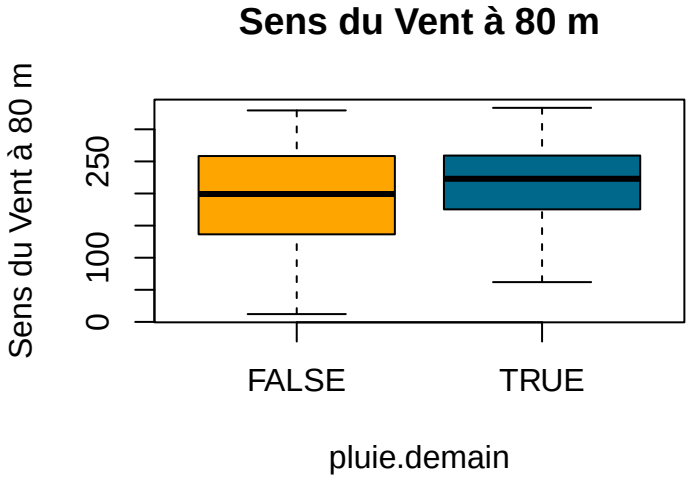
- L'impact de ***pluie.demain*** sur la distribution des covariables est **réel mais peu marqué**
  - les valeurs ***mean/min/max*** plus grandes **augmentent** le risque de pluie
  - les valeurs d'***amplitude*** plus grandes **augmentent** le risque de pluie
- Corrélations
  - ***mean/min/max*** sont **fortement corrélées** entre elles
  - l'***amplitude*** est fortement corrélée avec ***mean/max***
  - l'***amplitude*** est faiblement corrélée avec ***min***
- Dans le cadre de la sélection de variables (stratégie 2), on retiendra le produit **wind.speed.amplitude.80 x wind.speed.min.80**

3.2.10.2 Sens du vent :

covariable considérée

```
var_main="Sens du Vent à 80 m"
var_total=dat.meteo.train$wind.dir.80
```

Distribution selon pluie.demain



Analyse

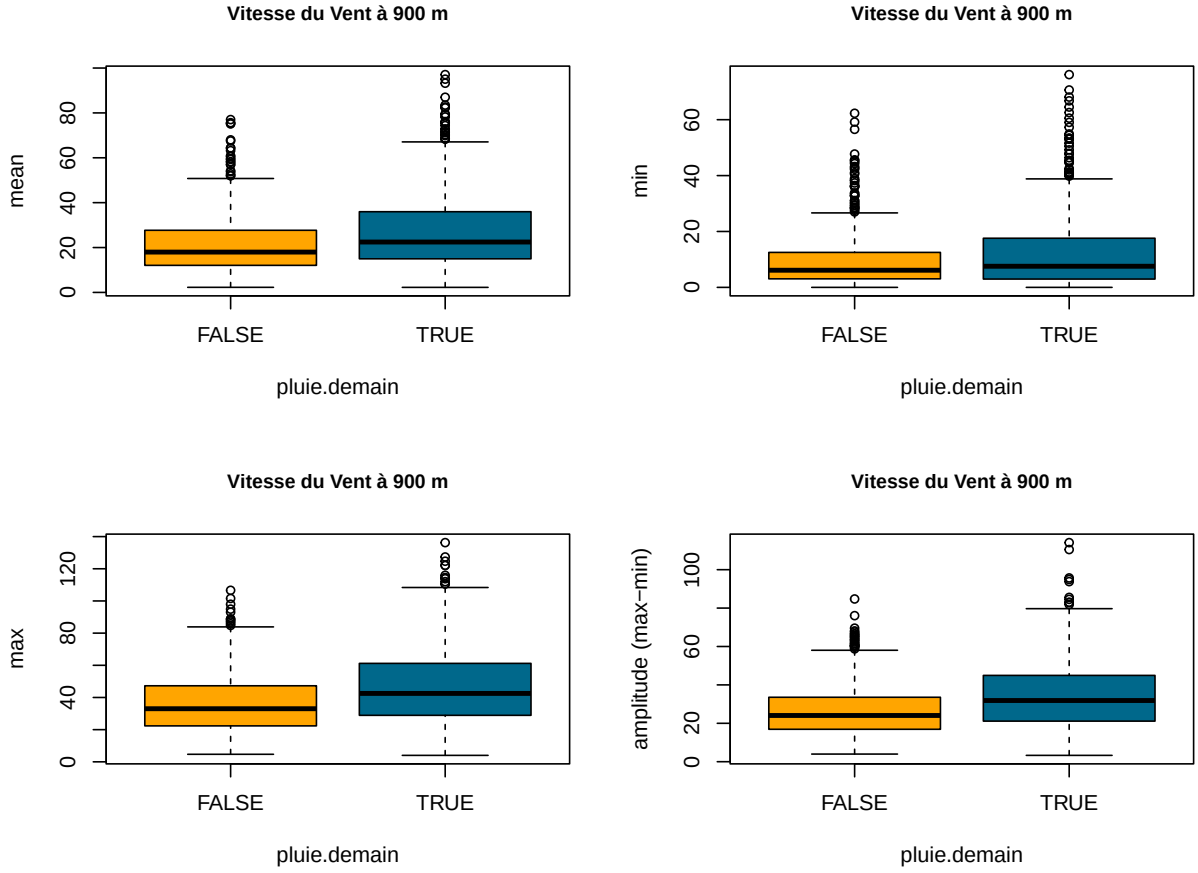
- L'impact de ***pluie.demain*** sur la distribution des covariables est **réel**
- Dans le cadre de la sélection de variables (stratégie 2), on retiendra le produit **wind.speed.amplitude.80 x wind.speed.min.80 x wind.dir.80**

3.2.11 Vitesse et sens du vent à 900 m

3.2.11.1 Vitesse du vent :

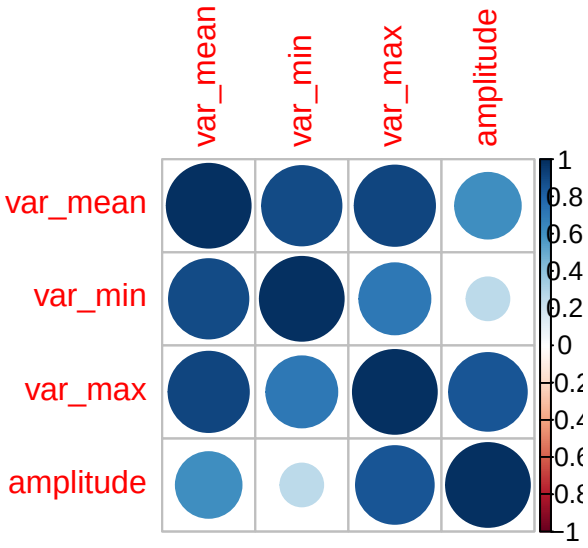
covariables considérées

Distribution des covariables selon pluie.demain



Corrélations entre covariables

```
##               var_mean   var_min   var_max amplitude
## var_mean   1.0000000 0.8953039 0.9168068 0.6133562
## var_min    0.8953039 1.0000000 0.7179283 0.2605610
## var_max    0.9168068 0.7179283 1.0000000 0.8591355
## amplitude 0.6133562 0.2605610 0.8591355 1.0000000
```



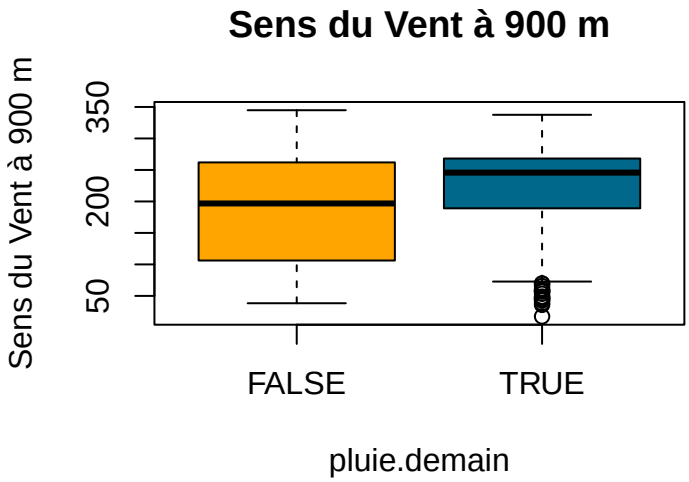
Analyse

- L'impact de ***pluie.demain*** sur la distribution des covariables est **réel mais peu marqué**
  - les valeurs ***mean/min/max*** plus grandes **augmentent** le risque de pluie
  - les valeurs d'***amplitude*** plus grandes **augmentent** le risque de pluie
- Corrélations
  - ***mean/min/max*** sont **fortement corrélées** entre elles
  - l'***amplitude*** est fortement corrélée avec ***mean/max***
  - l'***amplitude*** est faiblement corrélée avec ***min***
- Dans le cadre de la sélection de variables (stratégie 2), on retiendra le produit **wind.speed.amplitude.900 x wind.speed.min.900**

3.2.11.2 Sens du vent :

covariable considérée

Distribution selon pluie.demain



Analyse

- L'impact de ***pluie.demain*** sur la distribution des covariables est **réel**
- Dans le cadre de la sélection de variables (stratégie 2), on retiendra le produit **wind.speed.amplitude.900 x wind.speed.min.900 x wind.dir.900**

3.2.12 Corrélation entre vitesses et sens du vent

```
##                     pluie.demain wind.speed.mean.10 wind.speed.mean.80
## pluie.demain           1.0000000          0.2118871          0.1970128
## wind.speed.mean.10     0.2118871          1.0000000          0.9816588
## wind.speed.mean.80     0.1970128          0.9816588          1.0000000
## wind.speed.mean.900    0.1858268          0.7905158          0.7985150
##                     wind.speed.mean.900
## pluie.demain                  0.1858268
## wind.speed.mean.10            0.7905158
## wind.speed.mean.80            0.7985150
## wind.speed.mean.900           1.0000000
```

```
##                    pluie.demain wind.speed.min.10 wind.speed.min.80
## pluie.demain          1.0000000         0.1677079         0.1341332
## wind.speed.min.10     0.1677079         1.0000000         0.9333039
## wind.speed.min.80     0.1341332         0.9333039         1.0000000
## wind.speed.min.900    0.1210325         0.6500372         0.6666382
##                    wind.speed.min.900
## pluie.demain                0.1210325
## wind.speed.min.10           0.6500372
## wind.speed.min.80           0.6666382
## wind.speed.min.900          1.0000000
```

```
##                    pluie.demain wind.speed.max.10 wind.speed.max.80
## pluie.demain          1.0000000         0.2483077         0.2444658
## wind.speed.max.10     0.2483077         1.0000000         0.9476338
## wind.speed.max.80     0.2444658         0.9476338         1.0000000
## wind.speed.max.900    0.2379114         0.7699367         0.7799064
##                    wind.speed.max.900
## pluie.demain                0.2379114
## wind.speed.max.10           0.7699367
## wind.speed.max.80           0.7799064
## wind.speed.max.900          1.0000000
```

```
##                          pluie.demain wind.speed.amplitude.10
## pluie.demain                1.0000000               0.2257354
## wind.speed.amplitude.10     0.2257354               1.0000000
## wind.speed.amplitude.80     0.2199786               0.8787860
## wind.speed.amplitude.900    0.2409870               0.6076737
##                          wind.speed.amplitude.80 wind.speed.amplitude.900
## pluie.demain                           0.2199786                0.2409870
## wind.speed.amplitude.10                0.8787860                0.6076737
## wind.speed.amplitude.80                1.0000000                0.6454319
## wind.speed.amplitude.900               0.6454319                1.0000000
```

Analyse

- Les corrélations entre les données aux ***altitudes 10 et 80m*** sont fortement corrélées.
- Les corrélations avec les données à l'altitude 900m sont moins prononcées mais restent fortes

On prendra le parti d'inclure les données à 80m et 900m dans le cadre de la sélection de variables (stratégie 2) soit:

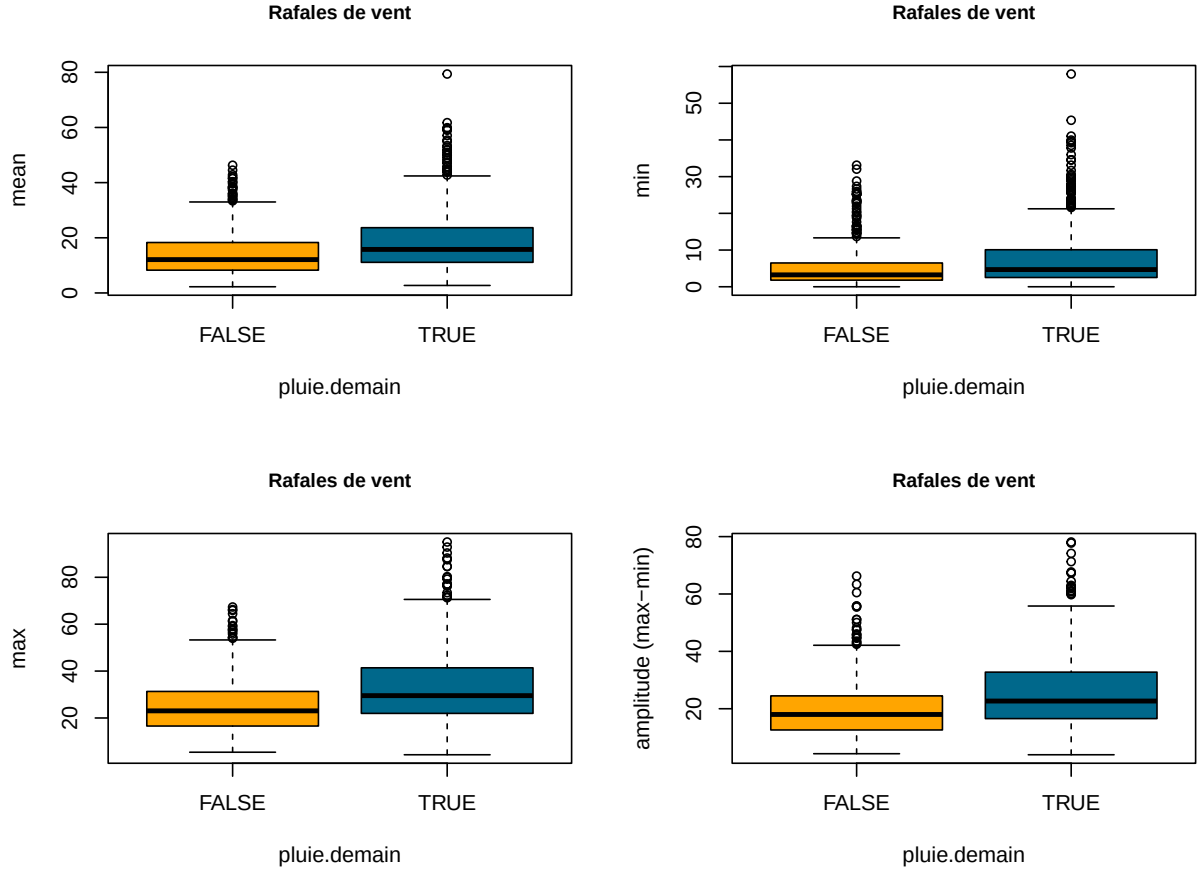
- **wind.speed.amplitude.900 x wind.speed.min.900 x wind.dir.900**
- **wind.speed.amplitude.80 x wind.speed.min.80 x wind.dir.80**

3.2.13 Rafales de vent

covariables considérées

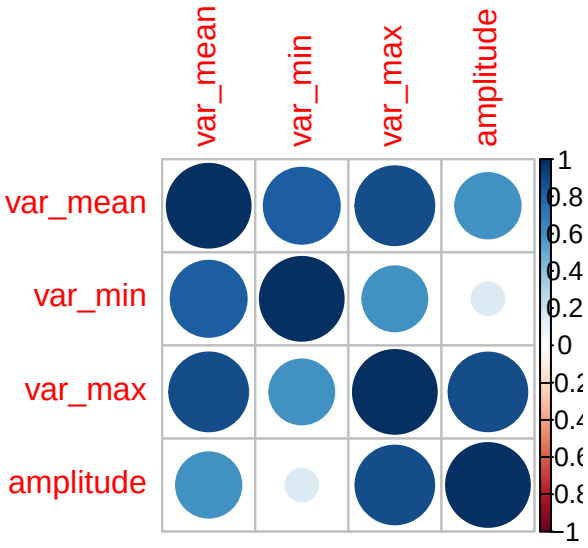
```
var_main="Rafales de vent"
var_mean=dat.meteo.train$wind.gust.mean
var_min=dat.meteo.train$wind.gust.min
var_max=dat.meteo.train$wind.gust.max
```

Distribution des covariables selon pluie.demain



Corrélations entre covariables

```
##               var_mean   var_min   var_max amplitude
## var_mean   1.0000000 0.8223337 0.8853670 0.6083197
## var_min    0.8223337 1.0000000 0.6007769 0.1513257
## var_max    0.8853670 0.6007769 1.0000000 0.8811236
## amplitude 0.6083197 0.1513257 0.8811236 1.0000000
```



Analyse

- L'impact de ***pluie.demain*** sur la distribution des covariables est **réel mais peu marqué**
  - les valeurs ***mean/min/max*** plus grandes **augmentent** le risque de pluie
  - les valeurs d'***amplitude*** plus grandes **augmentent** le risque de pluie
- Corrélations
  - ***mean/min/max*** sont **fortement corrélées** entre elles
  - l'***amplitude*** est fortement corrélée avec ***mean/max***
  - l'***amplitude*** est faiblement corrélée avec ***min***
- Dans le cadre de la sélection de variables (stratégie 2), on retiendra le produit **wind.gust.amplitude x wind.gust.min**

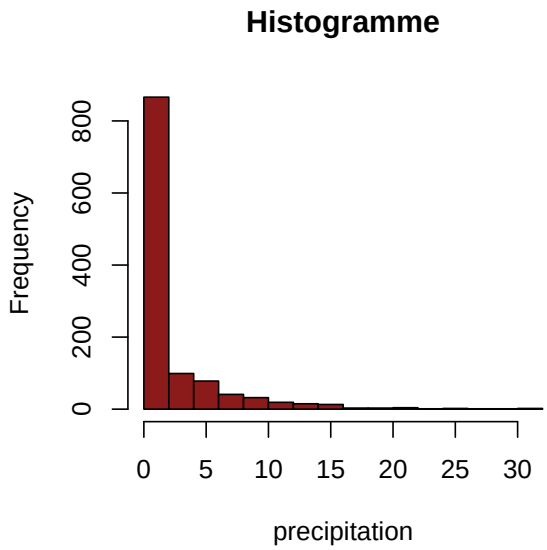
3.3 Analyse des covariables simples

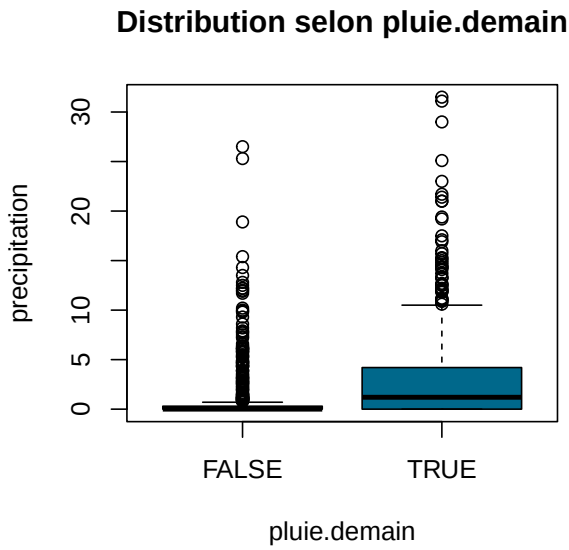
3.3.1 Précipitations

covariable considérée

```
var_main="Précipitations"
var_total=dat.meteo.train$precipitation
```

Histogramme et distribution selon pluie.demain





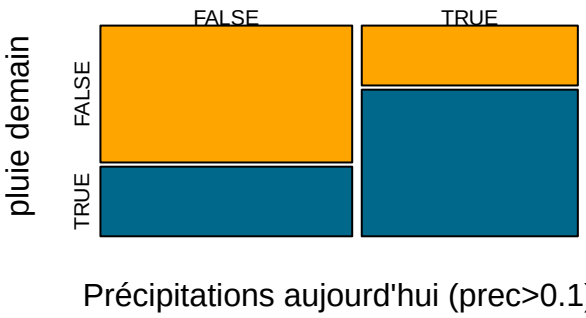
Quantiles de précipitations lorsque pluie.demain=FALSE

```
##   0%  25%  50%  75% 100%
##  0.0  0.0  0.0  0.3 26.5
```

Quantiles de précipitations lorsque pluie.demain=TRUE

```
##   0%  25%  50%  75% 100%
##  0.0  0.0  1.2  4.2 31.5
```

Distribution de la variable booléenne (précipitation $>$ 0.1) selon pluie.demain



Analyse

- Il pleut peu à Bâle et un grand nombre de valeurs sont nulles ou proches de 0.
- On définit arbitrairement un **seuil de 0.1** pour indiquer qu'il a plu et on crée une variable booléenne ***precipitation_bool*** (precipitation > 0.1).

On note :

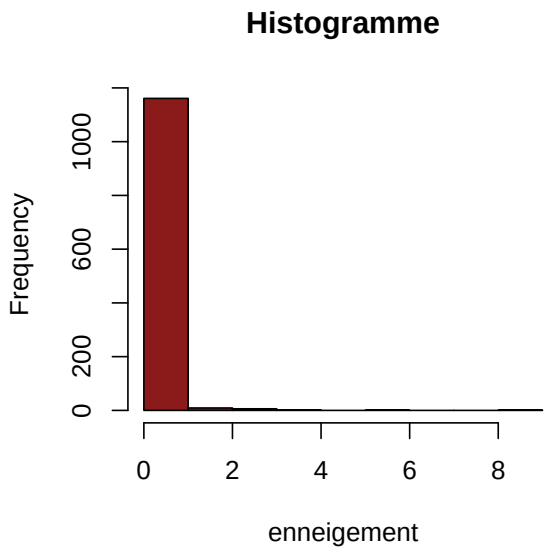
- Quand il a plu (**precipitation_bool** vaut ***TRUE***), le risque de pluie le lendemain est plus fort
- Quand il n'a pas plu (**precipitation_bool** vaut ***FALSE***), le risque de pluie le lendemain est plus faible
- On prendra le parti d'inclure les covariables **precipitation** et **precipitation_bool** dans le cadre de la sélection de variables (stratégie 2)

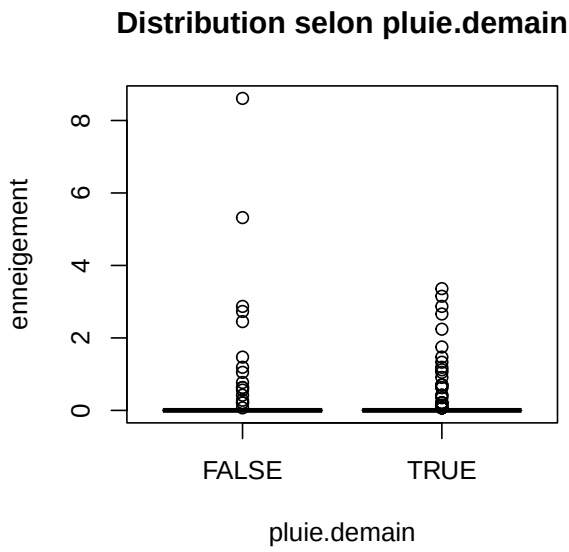
3.3.2 Enneigement

covariable considérée

```
var_main="Enneigement"
var_total=dat.meteo.train$snowfall
```

Histogramme et distribution selon pluie.demain





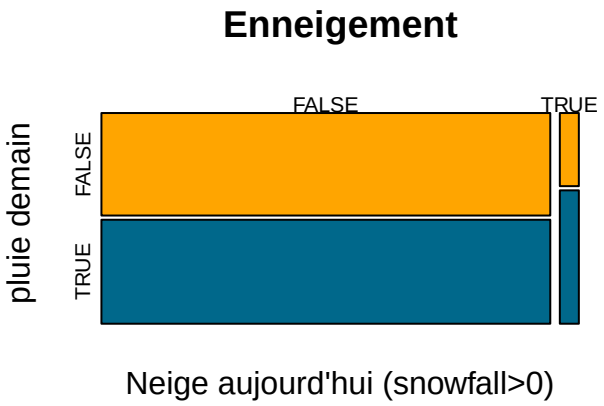
Quantiles d'enneigement lorsque pluie.demain=FALSE

```
##   0%  25%  50%  75% 100%
## 0.00 0.00 0.00 0.00 8.61
```

Quantiles d'enneigement lorsque pluie.demain=TRUE

```
##   0%  25%  50%  75% 100%
## 0.00 0.00 0.00 0.00 3.36
```

Distribution de la variable booléenne (snowfall $>$ 0) selon pluie.demain



Analyse

- Il neige peu à Bâle et un grand nombre de jour présente un enneigement nul.
- On définit une variable booléenne ***snowfall_bool*** (***snowfall*** > 0) pour indiquer qu'il a neigé.

On note que, quand il a neigé (**snowfall_bool** vaut ***TRUE***), le risque de pluie le lendemain est plus fort

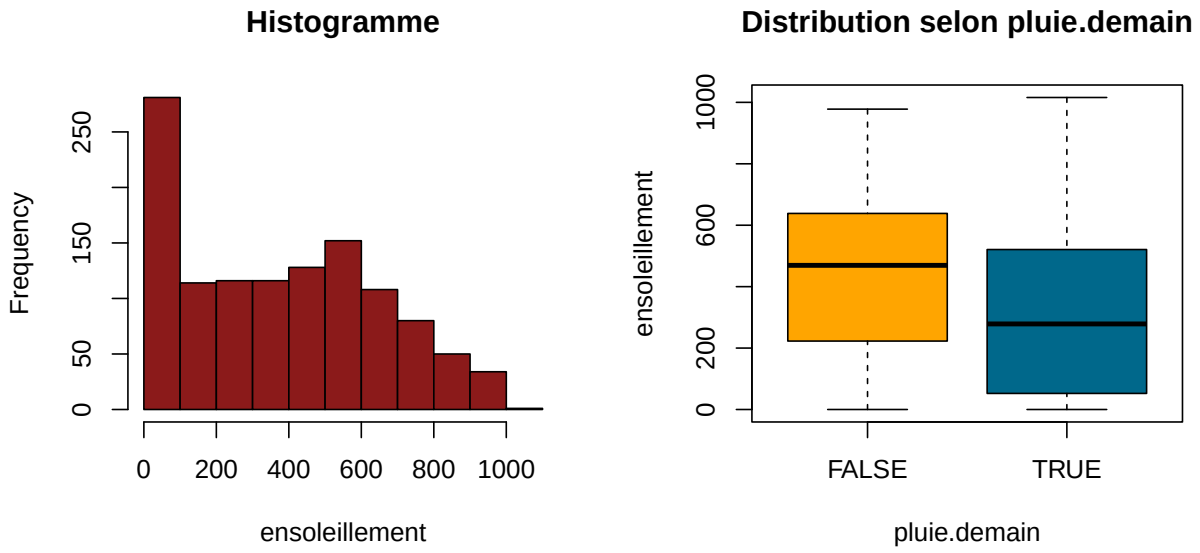
- On prendra le parti d'inclure les covariables **snowfall** et **snowfall_bool** dans le cadre de la sélection de variables (stratégie 2)

3.3.3 Ensoleillement

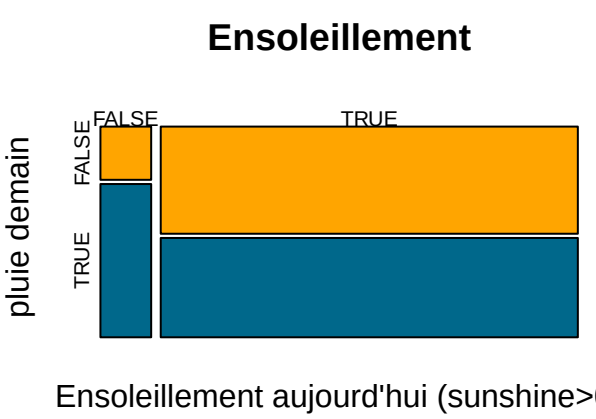
covariable considérée

```
var_main="Ensoleillement"
var_total=dat.meteo.train$sunshine
```

Histogramme et distribution selon pluie.demain



Distribution de la variable booléenne (sunshine>0) selon pluie.demain



Analyse

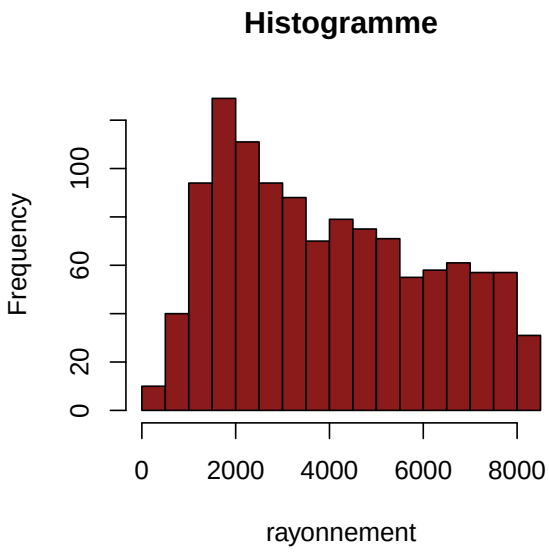
- On définit une variable booléenne ***sunshine_bool*** (***sunshine*** > 0) pour indiquer qu'il y a eu du soleil. On note que, quand il y a eu du soleil (**sunshine_bool** vaut ***TRUE***), le risque de pluie le lendemain est moins élevé.
- On prendra le parti d'inclure les covariables **sunshine** et **sunshine_bool** dans le cadre de la sélection de variables (stratégie 2)

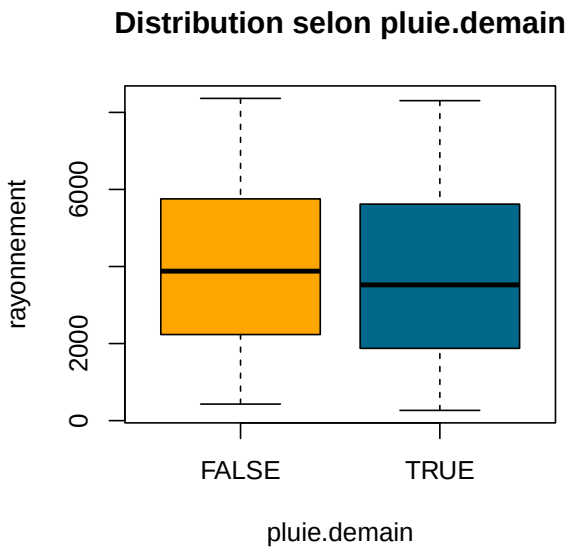
3.3.4 Rayonnement

covariable considérée

```
var_main="Rayonnement"
var_total=dat.meteo.train$radiation
```

Histogramme et distribution selon pluie.demain





Analyse

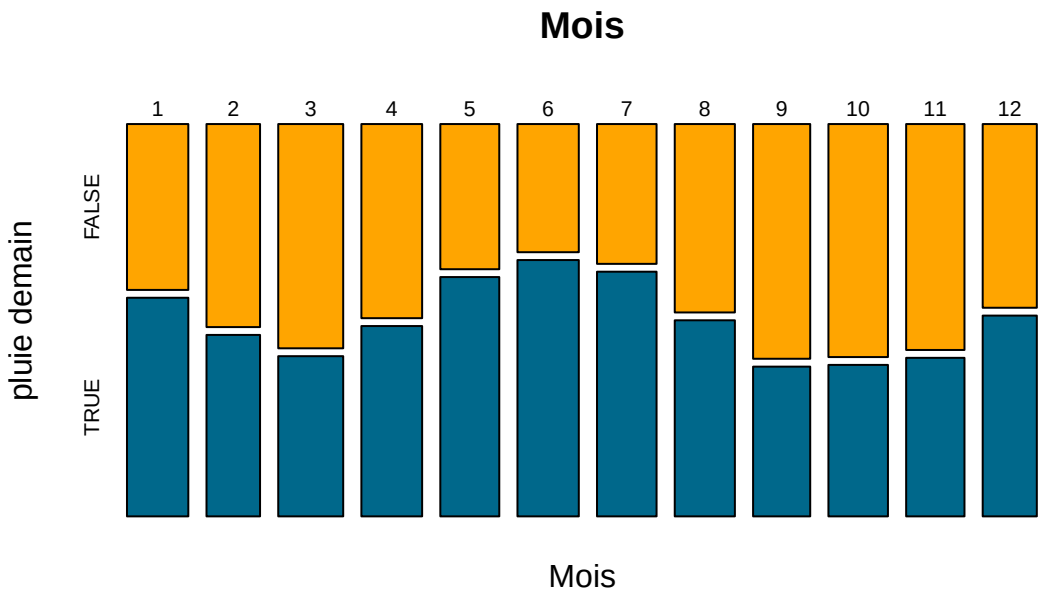
- L'impact de ***pluie.demain*** sur la distribution de **radiation** est **réel mais peu marqué**
- Dans le cadre de la sélection de variables (stratégie 2), on retiendra la covariable ***radiation***

3.3.5 Mois (catégorielle)

covariable considérée

```
var_main="Mois"
```

Distribution selon pluie.demain

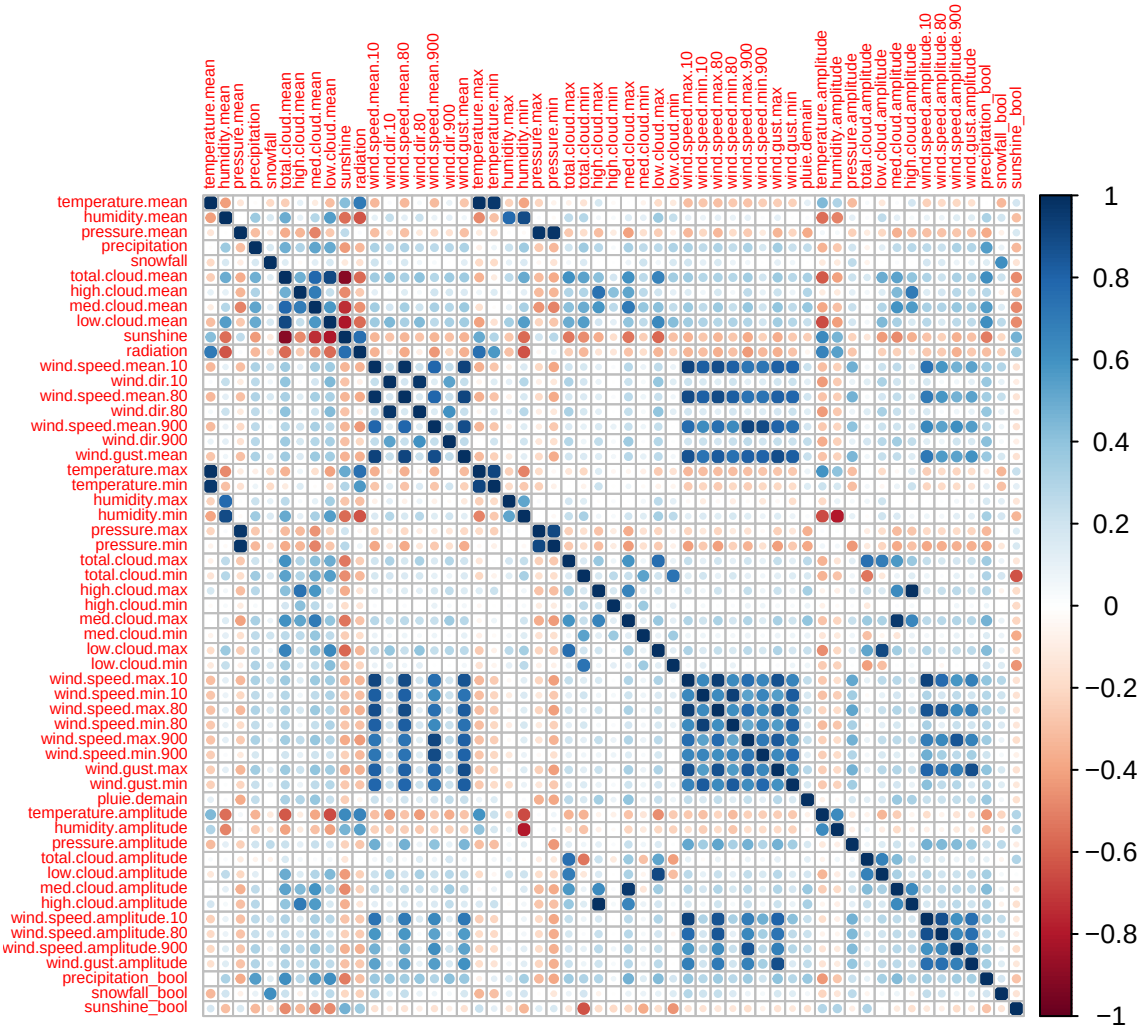


Analyse

- Le mois de l'année a une influence sur le risque de pluie le lendemain
- On inclura la covariable **month** sous **forme catégorielle** dans le cadre de la sélection de variables (stratégie 2).

3.4 Colinéarité entre familles de covariables - corrélation avec pluie.demain

Corrélation des covariables sur le jeu de données meteo.train



Dans chaque famille de covariables, on retrouve les schémas de corrélation mis en avant dans les chapitres précédemment.

On note en plus de fortes corrélations entre familles de covariables comme :

- entre ***radiation*** et ***sunshine***
- entre ***radiation*** et ***sunshine*** et la ***température***, les ***nébulosités*** à l'***humidité***
- l'***humidité*** minimale et la ***température*** moyenne.

On note aussi que la variable d'intérêt ***pluie.demain*** ne présente pas de corrélations fortement marquées.

Elle est principalement et raisonnablement corrélée :

- à la **pression** (covariables **pressure**)
- aux **nébulosités** (covariables ***cloud***)
- à la **précipitation** (notamment la covariable booléenne ***precipitation_bool***)
- dans une moindre mesure à la **vitesse des vents** (covariables ***wind.speed***)

Ce sont ces données qu'on s'attend à voir ressortir lors de l'ajustement des modèles.

3.5 Synthèse

L'analyse exploratoire a permis

- de comprendre les **schémas de corrélation** au sein des familles de covariables
- de créer des nouvelles covariables synthétiques (comme les amplitudes) ou booléennes pour simplifier la corrélation avec la variable d'intérêt ***pluie.demain***
- d'identifier les covariables les plus à même de représenter les familles

A partir de là, il est possible de construire un **modèle complet basé sur l'analyse exploratoire** contenant les covariables estimées être les plus pertinentes à savoir :

```
s2.res.glm.0.formula <- formula (

      pluie.demain ~

      # factors
      month +

      # single-values with mean/min/max/amplitude
      temperature.amplitude*temperature.min +
      humidity.amplitude*humidity.max +
      pressure.amplitude*pressure.max +

      # cloud
      total.cloud.amplitude*total.cloud.mean +
      low.cloud.amplitude*low.cloud.min +
      med.cloud.amplitude*med.cloud.min +
      high.cloud.amplitude*high.cloud.min +

      # wind
      wind.speed.amplitude.80*wind.speed.min.80*wind.dir.80 +
      wind.speed.amplitude.900*wind.speed.min.900*wind.dir.900 +
      wind.gust.amplitude*wind.gust.min +

      # others
      precipitation +
      precipitation_bool +
      snowfall +
      snowfall_bool +
      sunshine +
      sunshine_bool +
      radiation)
```

Ce sera la base d'une stratégie d'identification des modèles candidats alternative (**stratégie n°2**) qui viendra en concurrence d'une **stratégie naïve** basée sur un modèle complet contenant toutes les covariables initiales non transformées du jeu de données **meteo.train** (**stratégie n°1**, hors ***Year***, ***Day***, ***Hour***, ***Minute***)

```
s2.res.glm.0.formula <- formula (pluie.demain ~ . -Year -Day -Hour - Minute)
```

4 Modélisation

4.1 Jeu d'entraînement et de validation

Afin d'identifier le modèle le plus apte à la prédiction du jeu de données **meteo.test**, on va séparer le jeu de données **meteo.train** en 2 jeux de données tirés aléatoirement selon un ratio 80 / 20:

- 80% du jeu de données **meteo.train** servira à l'ajustement des modèles : ce sera le **jeu d'entraînement**
- 20% du jeu de données **meteo.train** servira à mesurer la capacité prédictive du modèle : ce sera le **jeu de validation**.

Pour déterminer les observations du jeu de données qui serviront à l'entraînement du modèle, on génère un vecteur **scp.train** de valeurs booléennes dont 80% valent ***TRUE*** et 20% valent ***FALSE***.

```
scp.train.size <- 0.8

scp.train = sample(c(TRUE, FALSE),
                   nrow(dat.meteo.train), replace=TRUE,
                   prob=c(scp.train.size, 1-scp.train.size))
```

A noter

Pour permettre la reproductibilité des résultats présentés dans ce rapport à chaque exécution du code, le vecteur est sauvegardé dans un fichier **ref.training__validation.rdata**. S'il est présent dans le répertoire d'exécution du script, le fichier est chargé et utilisé. Si le fichier n'est pas présent un nouveau tirage aléatoire est effectué et le résultat est sauvegardé dans le fichier **ref.training__validation.rdata**.

4.2 Stratégie 1 : approche naïve

L'idée de l'approche est d'appliquer une sélection des covariables à partir d'un modèle complet composé de toutes les covariables contenues dans le jeu de données **meteo.train**, qu'on aura au préalable traité de la façon suivante :

- les colonnes **Year**, **Day**, **Hour**, **Minute** sont supprimées
- la colonne **Month** a été transformée en covariable catégorielle

Le modèle complet initial sera donc constitué de **42 covariables**.

4.2.1 modèle complet (toutes variables initiales)

```
s1.res.glm.0.formula <- formula("pluie.demain ~ .")

s1.res.glm.0 <- glm(s1.res.glm.0.formula,
                    data=dat.meteo.train[scp.train,1:42],
                    family="binomial")
```

L'ajustement présente les résultats suivants :

```
##
## Call:
## glm(formula = s1.res.glm.0.formula, family = "binomial", data = dat.meteo.train[scp.train,
##     1:42])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4979  -0.8003   0.2445   0.7955   2.7847
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)          8.051e+01  1.449e+01   5.556 2.77e-08 ***
## month2              -4.776e-01  4.353e-01  -1.097  0.27261
## month3              -9.637e-01  4.662e-01  -2.067  0.03871 *
## month4              -7.251e-01  5.460e-01  -1.328  0.18414
## month5              -2.835e-01  6.197e-01  -0.458  0.64728
## month6               1.997e-01  7.087e-01   0.282  0.77809
## month7              -5.952e-02  7.126e-01  -0.084  0.93343
## month8              -7.161e-01  6.428e-01  -1.114  0.26530
## month9              -1.198e+00  5.605e-01  -2.137  0.03258 *
## month10             -9.694e-01  4.595e-01  -2.110  0.03489 *
## month11             -9.126e-01  4.390e-01  -2.079  0.03764 *
## month12             -9.374e-02  4.407e-01  -0.213  0.83157
## temperature.mean    -1.291e-02  1.887e-01  -0.068  0.94545
## humidity.mean        1.741e-02  3.689e-02   0.472  0.63705
## pressure.mean        3.676e-01  1.601e-01   2.296  0.02166 *
## precipitation        1.242e-02  3.235e-02   0.384  0.70109
## snowfall            -3.098e-01  2.597e-01  -1.193  0.23298
## total.cloud.mean     1.242e-02  1.378e-02   0.902  0.36726
## high.cloud.mean      2.778e-03  7.689e-03   0.361  0.71793
## med.cloud.mean      -1.941e-03  7.581e-03  -0.256  0.79793
## low.cloud.mean      -4.392e-03  9.168e-03  -0.479  0.63192
## sunshine             4.482e-04  1.065e-03   0.421  0.67383
## radiation           -9.721e-05  1.462e-04  -0.665  0.50610
## wind.speed.mean.10  -7.754e-02  1.149e-01  -0.675  0.49987
## wind.dir.10          8.366e-03  6.556e-03   1.276  0.20190
## wind.speed.mean.80  -7.046e-02  8.079e-02  -0.872  0.38310
## wind.dir.80         -9.582e-03  6.783e-03  -1.413  0.15777
## wind.speed.mean.900  4.653e-03  2.902e-02   0.160  0.87263
## wind.dir.900         4.532e-03  1.696e-03   2.672  0.00755 **
## wind.gust.mean       5.431e-02  4.127e-02   1.316  0.18820
## temperature.max      1.324e-01  1.104e-01   1.199  0.23042
## temperature.min     -7.203e-02  9.993e-02  -0.721  0.47103
## humidity.max        -2.341e-03  2.330e-02  -0.100  0.91997
## humidity.min        -1.222e-02  2.063e-02  -0.593  0.55345
## pressure.max        -1.907e-01  8.626e-02  -2.211  0.02703 *
## pressure.min        -2.599e-01  8.696e-02  -2.989  0.00280 **
## total.cloud.max      3.978e-03  5.807e-03   0.685  0.49330
## total.cloud.min      8.621e-03  7.461e-03   1.156  0.24788
## high.cloud.max       2.526e-04  3.331e-03   0.076  0.93954
## high.cloud.min      -1.664e-03  2.271e-02  -0.073  0.94161
## med.cloud.max        1.045e-02  3.646e-03   2.867  0.00414 **
## med.cloud.min        4.985e-03  1.063e-02   0.469  0.63919
## low.cloud.max        2.494e-03  3.951e-03   0.631  0.52792
## low.cloud.min       -3.854e-03  8.145e-03  -0.473  0.63607
## wind.speed.max.10    4.568e-02  4.053e-02   1.127  0.25970
## wind.speed.min.10    1.900e-01  7.264e-02   2.615  0.00892 **
## wind.speed.max.80   -3.018e-03  3.301e-02  -0.091  0.92716
## wind.speed.min.80   -9.207e-02  4.875e-02  -1.889  0.05895 .
## wind.speed.max.900  -1.213e-02  1.349e-02  -0.899  0.36861
## wind.speed.min.900   1.179e-02  2.204e-02   0.535  0.59274
## wind.gust.max        9.914e-03  1.895e-02   0.523  0.60083
## wind.gust.min        6.767e-03  3.162e-02   0.214  0.83053
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1296.85  on 935  degrees of freedom
## Residual deviance:  951.49  on 884  degrees of freedom
## AIC: 1055.5
##
## Number of Fisher Scoring iterations: 5
```

Le modèle est peu concluant ; peu de covariables sont significatives.

Les covariables les plus significatives sont:

- la **pression atmosphérique**
- la **nébulosité moyenne**
- la **direction du vent à 900m**
- Le **mois** dont certaines modalités sont significatives

On retrouve dans ce résultat les variables les plus corrélées à la variable d'intérêt **pluie.demain**.

Comme aucune précaution n'a été prise dans ce modèle pour limiter les colinéarités, il est probable que certaines covariables aient été négligées lors de l'ajustement.

Afin de sélectionner les covariables du modèle, on va procéder à une sélection avec les différentes méthodes de sélection pas-à-pas.

4.2.2 méthode ascendante

On procède à une sélection de covariables par **méthode pas-à-pas ascendante** (step forward)

- en partant d'un modèle constant **pluie.demain ~ 1**

```
# modèle constant initial
s1.model.constant <- glm (pluie.demain~1,
                          data=dat.meteo.train[scp.train,1:42],
                          family=binomial)
```

- vers le modèle complet

```
s1.model.full <- formula(s1.res.glm.0)
```

```
# Sélection des covariables par step forward
s1.res.step_forward <- step ( s1.model.constant,
                              scope=s1.model.full,
                              direction="forward")
```

Le modèle identifié par la **méthode pas-à-pas ascendante** est le suivant

```
## pluie.demain ~ med.cloud.max + pressure.min + month + wind.dir.900 +
##     temperature.max + total.cloud.mean + wind.gust.max + snowfall
```

Son ajustement présente les caractéristiques suivantes

```
##
## Call:
## glm(formula = pluie.demain ~ med.cloud.max + pressure.min + month +
##     wind.dir.900 + temperature.max + total.cloud.mean + wind.gust.max +
##     snowfall, family = binomial, data = dat.meteo.train[scp.train,
##     1:42])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3679  -0.8271   0.3558   0.8172   2.6499
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      75.223828  12.805093   5.875 4.24e-09 ***
## med.cloud.max     0.011131   0.002506   4.442 8.92e-06 ***
```

```
## pressure.min     -0.076827   0.012466  -6.163 7.13e-10 ***
## month2           -0.455023   0.413231  -1.101 0.270838
## month3           -1.081805   0.403506  -2.681 0.007340 **
## month4           -0.913793   0.429323  -2.128 0.033300 *
## month5           -0.662701   0.470370  -1.409 0.158867
## month6           -0.387512   0.498625  -0.777 0.437064
## month7           -0.551243   0.529624  -1.041 0.297959
## month8           -1.205932   0.514835  -2.342 0.019162 *
## month9           -1.548534   0.478339  -3.237 0.001207 **
## month10          -1.169700   0.425925  -2.746 0.006028 **
## month11          -1.106522   0.416834  -2.655 0.007941 **
## month12          -0.118143   0.417855  -0.283 0.777378
## wind.dir.900      0.002246   0.001172   1.917 0.055278 .
## temperature.max   0.076188   0.021838   3.489 0.000485 ***
## total.cloud.mean  0.010762   0.003660   2.940 0.003281 **
## wind.gust.max     0.014585   0.006666   2.188 0.028658 *
## snowfall         -0.374218   0.240664  -1.555 0.119961
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1296.85  on 935  degrees of freedom
## Residual deviance:  981.95  on 917  degrees of freedom
## AIC: 1020
##
## Number of Fisher Scoring iterations: 4
```

soit :

- 8 covariables
- 19 coefficients
- AIC = 1020

L'ajustement met en avant les covariables suivantes

- la **pression atmosphérique**
- la **nébulosité moyenne et totale**
- la **température maximale**
- la **direction du vent à 900m** ainsi que les **rafales**
- Le **mois** dont certaines modalités sont significatives
- l'**enneigement**

4.2.3 méthode progressive ascendante

On procède à une sélection de covariables par **méthode pas-à-pas progressive ascendante** (step both)

- en partant d'un modèle constant **pluie.demain ~ 1**

```
# modèle constant initial
s1.model.constant <- glm (pluie.demain~1,
                          data=dat.meteo.train[scp.train,1:42],
                          family=binomial)
```

- vers le modèle complet

```
s1.model.full <- formula(s1.res.glm.0)
```

```
s1.res.step_both_from_constant <- step (s1.model.constant,
                                        scope=s1.model.full,
                                        direction="both")
```

Le modèle identifié par la **méthode pas-à-pas progressive ascendante** est le suivant

```
## pluie.demain ~ med.cloud.max + pressure.min + month + wind.dir.900 +
##     temperature.max + total.cloud.mean + wind.gust.max + snowfall
```

Son ajustement présente les caractéristiques suivantes :

```
##
## Call:
## glm(formula = pluie.demain ~ med.cloud.max + pressure.min + month +
##     wind.dir.900 + temperature.max + total.cloud.mean + wind.gust.max +
##     snowfall, family = binomial, data = dat.meteo.train[scp.train,
##     1:42])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3679  -0.8271   0.3558   0.8172   2.6499
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      75.223828  12.805093   5.875 4.24e-09 ***
## med.cloud.max     0.011131   0.002506   4.442 8.92e-06 ***
## pressure.min     -0.076827   0.012466  -6.163 7.13e-10 ***
## month2           -0.455023   0.413231  -1.101 0.270838
## month3           -1.081805   0.403506  -2.681 0.007340 **
## month4           -0.913793   0.429323  -2.128 0.033300 *
## month5           -0.662701   0.470370  -1.409 0.158867
## month6           -0.387512   0.498625  -0.777 0.437064
## month7           -0.551243   0.529624  -1.041 0.297959
## month8           -1.205932   0.514835  -2.342 0.019162 *
## month9           -1.548534   0.478339  -3.237 0.001207 **
## month10          -1.169700   0.425925  -2.746 0.006028 **
## month11          -1.106522   0.416834  -2.655 0.007941 **
## month12          -0.118143   0.417855  -0.283 0.777378
## wind.dir.900      0.002246   0.001172   1.917 0.055278 .
## temperature.max   0.076188   0.021838   3.489 0.000485 ***
## total.cloud.mean  0.010762   0.003660   2.940 0.003281 **
## wind.gust.max     0.014585   0.006666   2.188 0.028658 *
## snowfall         -0.374218   0.240664  -1.555 0.119961
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1296.85  on 935  degrees of freedom
## Residual deviance:  981.95  on 917  degrees of freedom
## AIC: 1020
##
## Number of Fisher Scoring iterations: 4
```

Le modèle identifié par la méthode progressive ascendante est identique à celui identifié par la méthode ascendante :

- 8 covariables
- 19 coefficients
- AIC = 1020

L'ajustement met en avant les covariables suivantes

- la **pression atmosphérique**
- la **nébulosité moyenne et totale**
- la **température maximale**
- la **direction du vent à 900m** ainsi que les **rafales**
- Le **mois** dont certaines modalités sont significatives
- l'**enneigement**

4.2.4 méthode descendante

On procède à une sélection de covariables par **méthode pas-à-pas descendante** (step backward) partant du modèle complet.

```
s1.res.step_backward <- step (s1.res.glm.0, direction="backward")
```

Le modèle identifié par la **méthode pas-à-pas descendante** est le suivant

```
## pluie.demain ~ month + pressure.mean + total.cloud.mean + wind.speed.mean.10 +
##     wind.dir.10 + wind.dir.80 + wind.dir.900 + wind.gust.mean +
##     temperature.max + temperature.min + pressure.max + pressure.min +
##     total.cloud.min + med.cloud.max + low.cloud.max + wind.speed.max.10 +
##     wind.speed.min.10 + wind.speed.min.80
```

Son ajustement présente les caractéristiques suivantes :

```
##
## Call:
## glm(formula = pluie.demain ~ month + pressure.mean + total.cloud.mean +
##     wind.speed.mean.10 + wind.dir.10 + wind.dir.80 + wind.dir.900 +
##     wind.gust.mean + temperature.max + temperature.min + pressure.max +
##     pressure.min + total.cloud.min + med.cloud.max + low.cloud.max +
##     wind.speed.max.10 + wind.speed.min.10 + wind.speed.min.80,
##     family = "binomial", data = dat.meteo.train[scp.train, 1:42])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3433  -0.8034   0.2739   0.8129   2.7421
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        77.667963  13.601621   5.710 1.13e-08 ***
## month2             -0.537834   0.420488  -1.279  0.20087
## month3             -1.022646   0.421921  -2.424  0.01536 *
## month4             -0.852094   0.450482  -1.892  0.05856 .
## month5             -0.476213   0.501855  -0.949  0.34267
## month6             -0.029489   0.555021  -0.053  0.95763
## month7             -0.263711   0.577045  -0.457  0.64767
## month8             -0.882770   0.559355  -1.578  0.11452
## month9             -1.276671   0.511630  -2.495  0.01259 *
## month10            -1.061297   0.440827  -2.408  0.01606 *
## month11            -0.942738   0.430892  -2.188  0.02868 *
## month12            -0.137542   0.433297  -0.317  0.75092
## pressure.mean       0.371596   0.148071   2.510  0.01209 *
## total.cloud.mean    0.007691   0.005121   1.502  0.13312
## wind.speed.mean.10 -0.176484   0.063410  -2.783  0.00538 **
## wind.dir.10         0.009041   0.006379   1.417  0.15636
## wind.dir.80        -0.010311   0.006631  -1.555  0.11998
## wind.dir.900        0.004472   0.001631   2.743  0.00610 **
## wind.gust.mean      0.065477   0.022997   2.847  0.00441 **
## temperature.max     0.117456   0.041715   2.816  0.00487 **
## temperature.min    -0.067075   0.045999  -1.458  0.14479
## pressure.max       -0.203629   0.080097  -2.542  0.01101 *
```

```
## pressure.min       -0.247779   0.080777  -3.067  0.00216 **
## total.cloud.min     0.006311   0.004443   1.420  0.15552
## med.cloud.max       0.011799   0.002597   4.543 5.54e-06 ***
## low.cloud.max       0.004379   0.003055   1.433  0.15179
## wind.speed.max.10   0.045953   0.029100   1.579  0.11430
## wind.speed.min.10   0.204187   0.065086   3.137  0.00171 **
## wind.speed.min.80  -0.093514   0.042544  -2.198  0.02794 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1296.85  on 935  degrees of freedom
## Residual deviance:  958.68  on 907  degrees of freedom
## AIC: 1016.7
##
## Number of Fisher Scoring iterations: 4
```

Le modèle identifié présente les caractéristiques suivantes

- 18 covariables
- 29 coefficients
- AIC = 1016.7

L'ajustement met en avant les covariables suivantes

- la **pression atmosphérique**
- la **nébulosité totale, moyenne et basse**
- la **température**
- la **direction et la vitesse du vent** ainsi que les **rafales**
- Le **mois** dont certaines modalités sont significatives

4.2.5 méthode progressive descendante

On procède à une sélection de covariables par **méthode pas-à-pas progressive descendante** (step both) partant du modèle complet.

```
s1.res.step_both_from_full <- step (s1.res.glm.0, direction="both")
```

Le modèle identifié par la **méthode pas-à-pas progressive descendante** est le suivant

```
## pluie.demain ~ month + pressure.mean + total.cloud.mean + wind.speed.mean.10 +
##     wind.dir.10 + wind.dir.80 + wind.dir.900 + wind.gust.mean +
##     temperature.max + temperature.min + pressure.max + pressure.min +
##     total.cloud.min + med.cloud.max + low.cloud.max + wind.speed.max.10 +
##     wind.speed.min.10 + wind.speed.min.80
```

Son ajustement présente les caractéristiques suivantes :

```
##
## Call:
## glm(formula = pluie.demain ~ month + pressure.mean + total.cloud.mean +
##     wind.speed.mean.10 + wind.dir.10 + wind.dir.80 + wind.dir.900 +
##     wind.gust.mean + temperature.max + temperature.min + pressure.max +
##     pressure.min + total.cloud.min + med.cloud.max + low.cloud.max +
##     wind.speed.max.10 + wind.speed.min.10 + wind.speed.min.80,
##     family = "binomial", data = dat.meteo.train[scp.train, 1:42])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3433  -0.8034   0.2739   0.8129   2.7421
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        77.667963  13.601621   5.710 1.13e-08 ***
## month2             -0.537834   0.420488  -1.279  0.20087
## month3             -1.022646   0.421921  -2.424  0.01536 *
## month4             -0.852094   0.450482  -1.892  0.05856 .
## month5             -0.476213   0.501855  -0.949  0.34267
## month6             -0.029489   0.555021  -0.053  0.95763
## month7             -0.263711   0.577045  -0.457  0.64767
## month8             -0.882770   0.559355  -1.578  0.11452
## month9             -1.276671   0.511630  -2.495  0.01259 *
## month10            -1.061297   0.440827  -2.408  0.01606 *
## month11            -0.942738   0.430892  -2.188  0.02868 *
## month12            -0.137542   0.433297  -0.317  0.75092
## pressure.mean       0.371596   0.148071   2.510  0.01209 *
## total.cloud.mean    0.007691   0.005121   1.502  0.13312
## wind.speed.mean.10 -0.176484   0.063410  -2.783  0.00538 **
## wind.dir.10         0.009041   0.006379   1.417  0.15636
## wind.dir.80        -0.010311   0.006631  -1.555  0.11998
## wind.dir.900        0.004472   0.001631   2.743  0.00610 **
## wind.gust.mean      0.065477   0.022997   2.847  0.00441 **
## temperature.max     0.117456   0.041715   2.816  0.00487 **
## temperature.min    -0.067075   0.045999  -1.458  0.14479
## pressure.max       -0.203629   0.080097  -2.542  0.01101 *
## pressure.min       -0.247779   0.080777  -3.067  0.00216 **
## total.cloud.min     0.006311   0.004443   1.420  0.15552
## med.cloud.max       0.011799   0.002597   4.543 5.54e-06 ***
## low.cloud.max       0.004379   0.003055   1.433  0.15179
## wind.speed.max.10   0.045953   0.029100   1.579  0.11430
## wind.speed.min.10   0.204187   0.065086   3.137  0.00171 **
## wind.speed.min.80  -0.093514   0.042544  -2.198  0.02794 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1296.85  on 935  degrees of freedom
## Residual deviance:  958.68  on 907  degrees of freedom
## AIC: 1016.7
##
## Number of Fisher Scoring iterations: 4
```

Le modèle identifié par la méthode progressive ascendante est identique à celui identifié par la méthode ascendante:

- 18 covariables
- 29 coefficients
- AIC = 1016.7

L'ajustement met en avant les covariables suivantes

- la **pression atmosphérique**
- la **nébulosité totale, moyenne et basse**
- la **température**
- la **direction et la vitesse du vent** ainsi que les **rafales**
- Le **mois** dont certaines modalités sont significatives

4.3 Stratégie 2 : approche par l'analyse exploratoire

L'idée de l'approche est d'initialiser la démarche d'identification de modèle par un modèle complet composé d'une sélection de covariables issues de l'analyse exploratoire (cf. chapitre **Analyse exploratoire** > **Synthèse**).

Pour coller au plus près des constatations effectuées, le jeu de données **meteo.train** a été enrichi de la façon suivante :

- Des variables **amplitude** (max-min) ont été ajoutées en fin de tableau pour toutes les variables sous la forme **mean/min/max**.
- Des **variables booléennes** ont été ajoutées en fin de tableau pour les données ***precipitation, snowfall et sunshine***.

Le jeu de données est donc constitué de **55 covariables**.

4.3.1 modèle complet (sélection de variables)

Le modèle complet initial est constitué de la façon suivante :

```
s2.res.glm.0 <- glm(pluie.demain ~

                    # factors
                    month +

                    # single-values with mean/min/max/amplitude
                    temperature.amplitude*temperature.min +
                    humidity.amplitude*humidity.max +
                    pressure.amplitude*pressure.max +

                    # cloud
                    total.cloud.amplitude*total.cloud.mean +
                    low.cloud.amplitude*low.cloud.min +
                    med.cloud.amplitude*med.cloud.min +
                    high.cloud.amplitude*high.cloud.min +

                    # wind
                    wind.speed.amplitude.80*wind.speed.min.80*wind.dir.80 +
                    wind.speed.amplitude.900*wind.speed.min.900*wind.dir.900 +
                    wind.gust.amplitude*wind.gust.min +

                    # others
                    precipitation +
                    precipitation_bool +
                    snowfall +
                    snowfall_bool +
                    sunshine +
                    sunshine_bool +
                    radiation,

                  data=dat.meteo.train[scp.train,],
                  family=binomial,
                  na.action=na.exclude)
```

L'ajustement présente les résultats suivants :

```
##
## Call:
## glm(formula = pluie.demain ~ month + temperature.amplitude *
##     temperature.min + humidity.amplitude * humidity.max + pressure.amplitude *
##     pressure.max + total.cloud.amplitude * total.cloud.mean +
##     low.cloud.amplitude * low.cloud.min + med.cloud.amplitude *
##     med.cloud.min + high.cloud.amplitude * high.cloud.min + wind.speed.amplitude.80 *
##     wind.speed.min.80 * wind.dir.80 + wind.speed.amplitude.900 *
##     wind.speed.min.900 * wind.dir.900 + wind.gust.amplitude *
##     wind.gust.min + precipitation + precipitation_bool + snowfall +
##     snowfall_bool + sunshine + sunshine_bool + radiation, family = binomial,
##     data = dat.meteo.train[scp.train, ], na.action = na.exclude)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4188  -0.8158   0.2959   0.8171   2.9125
##
## Coefficients:
##                                                          Estimate Std. Error
## (Intercept)                                             9.133e+01  2.637e+01
## month2                                                 -2.349e-01  4.448e-01
## month3                                                 -8.571e-01  5.007e-01
## month4                                                 -6.752e-01  5.878e-01
## month5                                                 -3.654e-01  6.583e-01
## month6                                                  1.857e-01  7.281e-01
## month7                                                 -1.764e-02  7.331e-01
## month8                                                 -6.784e-01  6.761e-01
## month9                                                 -1.081e+00  5.931e-01
## month10                                                -7.752e-01  4.982e-01
## month11                                                -7.340e-01  4.517e-01
## month12                                                 1.793e-03  4.418e-01
## temperature.amplitude                                   4.888e-02  6.691e-02
## temperature.min                                        -1.246e-02  5.970e-02
## humidity.amplitude                                     -2.118e-02  1.068e-01
## humidity.max                                           -1.407e-02  4.194e-02
## pressure.amplitude                                     -2.070e+00  3.623e+00
## pressure.max                                           -9.005e-02  2.541e-02
## total.cloud.amplitude                                   6.869e-03  5.760e-03
## total.cloud.mean                                        1.906e-02  1.257e-02
## low.cloud.amplitude                                     1.781e-03  3.933e-03
## low.cloud.min                                          -1.534e-03  9.206e-03
## med.cloud.amplitude                                     9.242e-03  3.382e-03
## med.cloud.min                                           2.662e-02  1.698e-02
## high.cloud.amplitude                                    2.634e-03  2.749e-03
## high.cloud.min                                          1.853e-02  4.507e-02
## wind.speed.amplitude.80                                -1.199e-02  6.603e-02
## wind.speed.min.80                                      -8.481e-02  2.035e-01
## wind.dir.80                                            -4.089e-03  6.157e-03
## wind.speed.amplitude.900                               -4.602e-02  3.495e-02
## wind.speed.min.900                                     -1.132e-01  8.165e-02
## wind.dir.900                                           -1.680e-03  4.383e-03
## wind.gust.amplitude                                     3.344e-02  1.718e-02
## wind.gust.min                                           7.361e-02  4.982e-02
## precipitation                                          -1.607e-02  3.116e-02
## precipitation_boolTRUE                                  4.407e-01  2.422e-01
## snowfall                                               -5.550e-01  3.779e-01
## snowfall_boolTRUE                                       3.509e-01  7.087e-01
## sunshine                                                3.152e-04  1.043e-03
## sunshine_boolTRUE                                      -3.386e-01  3.528e-01
## radiation                                              -4.471e-05  1.358e-04
## temperature.amplitude:temperature.min                   6.341e-03  5.368e-03
## humidity.amplitude:humidity.max                         3.617e-04  1.193e-03
## pressure.amplitude:pressure.max                         2.099e-03  3.547e-03
## total.cloud.amplitude:total.cloud.mean                 -1.236e-04  9.845e-05
## low.cloud.amplitude:low.cloud.min                      -1.106e-05  2.492e-04
## med.cloud.amplitude:med.cloud.min                      -3.254e-04  3.919e-04
## high.cloud.amplitude:high.cloud.min                    -1.516e-04  6.407e-04
## wind.speed.amplitude.80:wind.speed.min.80              -7.693e-03  1.006e-02
## wind.speed.amplitude.80:wind.dir.80                    -5.474e-05  2.968e-04
## wind.speed.min.80:wind.dir.80                           2.528e-05  8.465e-04
## wind.speed.amplitude.900:wind.speed.min.900             4.161e-03  2.662e-03
## wind.speed.amplitude.900:wind.dir.900                   1.751e-04  1.513e-04
## wind.speed.min.900:wind.dir.900                         4.863e-04  3.396e-04
## wind.gust.amplitude:wind.gust.min                      -7.120e-04  1.730e-03
## wind.speed.amplitude.80:wind.speed.min.80:wind.dir.80   4.019e-05  4.208e-05
## wind.speed.amplitude.900:wind.speed.min.900:wind.dir.900 -1.790e-05  1.092e-05
##                                                         z value Pr(>|z|)
## (Intercept)                                               3.464 0.000533 ***
```

```
## month2                                                -0.528 0.597395
## month3                                                -1.712 0.086943 .
## month4                                                -1.149 0.250729
## month5                                                -0.555 0.578876
## month6                                                 0.255 0.798647
## month7                                                -0.024 0.980807
## month8                                                -1.003 0.315625
## month9                                                -1.823 0.068253 .
## month10                                               -1.556 0.119728
## month11                                               -1.625 0.104166
## month12                                                0.004 0.996762
## temperature.amplitude                                  0.731 0.465037
## temperature.min                                       -0.209 0.834606
## humidity.amplitude                                    -0.198 0.842699
## humidity.max                                          -0.336 0.737215
## pressure.amplitude                                    -0.571 0.567776
## pressure.max                                          -3.543 0.000395 ***
## total.cloud.amplitude                                  1.193 0.233040
## total.cloud.mean                                       1.516 0.129421
## low.cloud.amplitude                                    0.453 0.650766
## low.cloud.min                                         -0.167 0.867661
## med.cloud.amplitude                                    2.733 0.006283 **
## med.cloud.min                                          1.568 0.116944
## high.cloud.amplitude                                   0.958 0.337975
## high.cloud.min                                         0.411 0.680985
## wind.speed.amplitude.80                               -0.182 0.855880
## wind.speed.min.80                                     -0.417 0.676843
## wind.dir.80                                           -0.664 0.506611
## wind.speed.amplitude.900                              -1.317 0.187930
## wind.speed.min.900                                    -1.386 0.165707
## wind.dir.900                                          -0.383 0.701553
## wind.gust.amplitude                                    1.947 0.051537 .
## wind.gust.min                                          1.477 0.139583
## precipitation                                         -0.516 0.605959
## precipitation_boolTRUE                                 1.819 0.068839 .
## snowfall                                              -1.469 0.141888
## snowfall_boolTRUE                                      0.495 0.620499
## sunshine                                               0.302 0.762539
## sunshine_boolTRUE                                     -0.960 0.337057
## radiation                                             -0.329 0.741910
## temperature.amplitude:temperature.min                  1.181 0.237494
## humidity.amplitude:humidity.max                        0.303 0.761735
## pressure.amplitude:pressure.max                        0.592 0.554056
## total.cloud.amplitude:total.cloud.mean                -1.256 0.209127
## low.cloud.amplitude:low.cloud.min                     -0.044 0.964588
## med.cloud.amplitude:med.cloud.min                     -0.830 0.406300
## high.cloud.amplitude:high.cloud.min                   -0.237 0.812900
## wind.speed.amplitude.80:wind.speed.min.80             -0.764 0.444589
## wind.speed.amplitude.80:wind.dir.80                   -0.184 0.853698
## wind.speed.min.80:wind.dir.80                          0.030 0.976179
## wind.speed.amplitude.900:wind.speed.min.900            1.563 0.118072
## wind.speed.amplitude.900:wind.dir.900                  1.157 0.247223
## wind.speed.min.900:wind.dir.900                        1.432 0.152202
## wind.gust.amplitude:wind.gust.min                     -0.412 0.680601
## wind.speed.amplitude.80:wind.speed.min.80:wind.dir.80  0.955 0.339531
## wind.speed.amplitude.900:wind.speed.min.900:wind.dir.900 -1.639 0.101276
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1296.85  on 935  degrees of freedom
## Residual deviance:  950.28  on 879  degrees of freedom
## AIC: 1064.3
##
## Number of Fisher Scoring iterations: 5
```

Le modèle est peu concluant ; peu de covariables sont significatives.

Les covariables mises en avant par le modèle sont:

- la **pression atmosphérique**
- la **nébulosité moyenne**

Afin de sélectionner les covariables du modèle, on va procéder à une sélection avec les différentes méthodes de sélection pas-à-pas.

4.3.2 méthode ascendante

On procède à une sélection de covariables par **méthode pas-à-pas ascendante** (step forward)

- en partant d'un modèle constant **pluie.demain ~ 1**

```
s2.model.constant <- glm (pluie.demain~1,
                          data=dat.meteo.train[scp.train,],
                          family=binomial)
```

- vers le modèle complet

```
s2.model.full <- formula(s2.res.glm.0)
```

```
s2.res.step_forward <- step (s2.model.constant,
                             scope=s2.model.full,
                             data=dat.meteo.train[scp.train,],
                             direction="forward")
```

Le modèle identifié par la **méthode pas-à-pas ascendante** est le suivant

```
## pluie.demain ~ med.cloud.amplitude + pressure.max + precipitation_bool +
##     month + med.cloud.min + pressure.amplitude + snowfall + temperature.amplitude +
##     wind.dir.900 + total.cloud.mean + temperature.min + wind.speed.min.900
```

Son ajustement :

```
##
## Call:
## glm(formula = pluie.demain ~ med.cloud.amplitude + pressure.max +
##     precipitation_bool + month + med.cloud.min + pressure.amplitude +
##     snowfall + temperature.amplitude + wind.dir.900 + total.cloud.mean +
##     temperature.min + wind.speed.min.900, family = binomial,
##     data = dat.meteo.train[scp.train, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3633  -0.8267   0.3313   0.8129   2.6764
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            76.533907  13.267332   5.769 7.99e-09 ***
## med.cloud.amplitude     0.011251   0.002534   4.440 9.01e-06 ***
## pressure.max           -0.078352   0.012891  -6.078 1.22e-09 ***
## precipitation_boolTRUE  0.410448   0.214949   1.910 0.056196 .
## month2                 -0.523744   0.416431  -1.258 0.208502
## month3                 -1.193949   0.408436  -2.923 0.003464 **
## month4                 -1.019397   0.436050  -2.338 0.019398 *
## month5                 -0.724654   0.477556  -1.517 0.129161
## month6                 -0.314351   0.520381  -0.604 0.545791
## month7                 -0.493971   0.551286  -0.896 0.370235
## month8                 -1.054699   0.537845  -1.961 0.049882 *
## month9                 -1.462492   0.492570  -2.969 0.002987 **
## month10                -1.163947   0.432595  -2.691 0.007132 **
## month11                -1.059749   0.423176  -2.504 0.012270 *
## month12                -0.049902   0.418910  -0.119 0.905178
## med.cloud.min           0.021805   0.009202   2.370 0.017803 *
## pressure.amplitude      0.076555   0.027387   2.795 0.005186 **
## snowfall               -0.466119   0.248476  -1.876 0.060668 .
## temperature.amplitude   0.132722   0.038100   3.484 0.000495 ***
## wind.dir.900            0.002555   0.001291   1.980 0.047726 *
## total.cloud.mean        0.010132   0.004403   2.301 0.021364 *
## temperature.min         0.051994   0.025691   2.024 0.042991 *
## wind.speed.min.900      0.012973   0.007740   1.676 0.093715 .
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1296.85  on 935  degrees of freedom
## Residual deviance:  976.12  on 913  degrees of freedom
## AIC: 1022.1
##
## Number of Fisher Scoring iterations: 4
```

Le modèle identifié présente les caractéristiques suivantes

- 12 covariables
- 23 coefficients
- AIC = 1022.1

L'ajustement met en avant les covariables suivantes

- la **pression atmosphérique**
- la **température**
- la **nébulosité moyenne** dont l'amplitude et le minima
- la **vitesse et direction du vent à 900m**
- Le **mois** dont certaines modalités sont significatives
- Les **précipitations** sous la forme booléenne
- l'**enneigement**

4.3.3 méthode progressive ascendante

On procède à une sélection de covariables par **méthode pas-à-pas progressive ascendante** (step both)

- en partant d'un modèle constant **pluie.demain ~ 1**

```
s2.model.constant <- glm (pluie.demain~1,
                          data=dat.meteo.train[scp.train,],
                          family=binomial)
```

- vers le modèle complet

```
s2.model.full <- formula(s2.res.glm.0)
```

```
s2.res.step_both_from_constant <- step (s2.model.constant,
                                        scope=s2.model.full,
                                        direction="both")
```

Le modèle identifié par la **méthode pas-à-pas progressive ascendante** est le suivant

```
## pluie.demain ~ med.cloud.amplitude + pressure.max + precipitation_bool +
##     month + med.cloud.min + pressure.amplitude + snowfall + temperature.amplitude +
##     wind.dir.900 + total.cloud.mean + temperature.min + wind.speed.min.900
```

Son ajustement :

```
##
## Call:
## glm(formula = pluie.demain ~ med.cloud.amplitude + pressure.max +
##     precipitation_bool + month + med.cloud.min + pressure.amplitude +
##     snowfall + temperature.amplitude + wind.dir.900 + total.cloud.mean +
##     temperature.min + wind.speed.min.900, family = binomial,
##     data = dat.meteo.train[scp.train, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3633  -0.8267   0.3313   0.8129   2.6764
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              76.533907  13.267332   5.769 7.99e-09 ***
## med.cloud.amplitude       0.011251   0.002534   4.440 9.01e-06 ***
## pressure.max             -0.078352   0.012891  -6.078 1.22e-09 ***
## precipitation_boolTRUE    0.410448   0.214949   1.910 0.056196 .
## month2                   -0.523744   0.416431  -1.258 0.208502
## month3                   -1.193949   0.408436  -2.923 0.003464 **
## month4                   -1.019397   0.436050  -2.338 0.019398 *
## month5                   -0.724654   0.477556  -1.517 0.129161
## month6                   -0.314351   0.520381  -0.604 0.545791
## month7                   -0.493971   0.551286  -0.896 0.370235
## month8                   -1.054699   0.537845  -1.961 0.049882 *
## month9                   -1.462492   0.492570  -2.969 0.002987 **
## month10                  -1.163947   0.432595  -2.691 0.007132 **
## month11                  -1.059749   0.423176  -2.504 0.012270 *
## month12                  -0.049902   0.418910  -0.119 0.905178
## med.cloud.min             0.021805   0.009202   2.370 0.017803 *
## pressure.amplitude        0.076555   0.027387   2.795 0.005186 **
## snowfall                 -0.466119   0.248476  -1.876 0.060668 .
## temperature.amplitude     0.132722   0.038100   3.484 0.000495 ***
## wind.dir.900              0.002555   0.001291   1.980 0.047726 *
## total.cloud.mean          0.010132   0.004403   2.301 0.021364 *
## temperature.min           0.051994   0.025691   2.024 0.042991 *
## wind.speed.min.900        0.012973   0.007740   1.676 0.093715 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1296.85  on 935  degrees of freedom
## Residual deviance:  976.12  on 913  degrees of freedom
## AIC: 1022.1
##
## Number of Fisher Scoring iterations: 4
```

Le modèle identifié par la méthode progressive ascendante est identique à celui identifié par la méthode ascendante.

Il présente les caractéristiques suivantes

- 12 covariables
- 23 coefficients
- AIC = 1022.1

L'ajustement met en avant les covariables suivantes

- la **pression atmosphérique**
- la **température**

- la **nébulosité moyenne** dont l'amplitude et le minima
- la **vitesse et direction du vent à 900m**
- Le **mois** dont certaines modalités sont significatives
- Les **précipitations** sous la forme booléenne
- l'**enneigement**

4.3.4 méthode descendante

On procède à une sélection de covariables par **méthode pas-à-pas descendante** (step backward) à partir du modèle complet.

```
s2.res.step_backward <- step (s2.res.glm.0,
                              direction="backward")
```

Le modèle identifié par la **méthode pas-à-pas descendante** est le suivant

```
## pluie.demain ~ temperature.amplitude + temperature.min + pressure.amplitude +
##     pressure.max + total.cloud.amplitude + total.cloud.mean +
##     med.cloud.amplitude + med.cloud.min + wind.speed.min.80 +
##     wind.dir.80 + wind.dir.900 + wind.gust.amplitude + wind.gust.min +
##     precipitation_bool + snowfall + temperature.amplitude:temperature.min +
##     total.cloud.amplitude:total.cloud.mean + wind.speed.min.80:wind.dir.80
```

Son ajustement :

```
##
## Call:
## glm(formula = pluie.demain ~ temperature.amplitude + temperature.min +
##     pressure.amplitude + pressure.max + total.cloud.amplitude +
##     total.cloud.mean + med.cloud.amplitude + med.cloud.min +
##     wind.speed.min.80 + wind.dir.80 + wind.dir.900 + wind.gust.amplitude +
##     wind.gust.min + precipitation_bool + snowfall + temperature.amplitude:temperature.min +
##     total.cloud.amplitude:total.cloud.mean + wind.speed.min.80:wind.dir.80,
##     family = binomial, data = dat.meteo.train[scp.train, ], na.action = na.exclude)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2652  -0.8478   0.3199   0.8290   2.8122
##
## Coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                             6.757e+01  1.298e+01   5.205 1.94e-07
## temperature.amplitude                   3.438e-03  4.897e-02   0.070 0.944031
## temperature.min                        -6.540e-02  4.061e-02  -1.610 0.107338
## pressure.amplitude                      6.002e-02  2.802e-02   2.142 0.032203
## pressure.max                           -6.908e-02  1.258e-02  -5.490 4.02e-08
## total.cloud.amplitude                   7.666e-03  4.466e-03   1.717 0.086064
## total.cloud.mean                        1.835e-02  6.630e-03   2.768 0.005637
## med.cloud.amplitude                     9.937e-03  2.652e-03   3.748 0.000179
## med.cloud.min                           1.606e-02  9.602e-03   1.673 0.094409
## wind.speed.min.80                      -1.873e-01  6.685e-02  -2.802 0.005075
## wind.dir.80                            -4.113e-03  2.104e-03  -1.955 0.050599
## wind.dir.900                            2.837e-03  1.445e-03   1.963 0.049663
## wind.gust.amplitude                     1.720e-02  7.638e-03   2.253 0.024289
## wind.gust.min                           5.140e-02  2.232e-02   2.303 0.021297
## precipitation_boolTRUE                  4.776e-01  2.135e-01   2.237 0.025286
## snowfall                               -4.514e-01  2.193e-01  -2.059 0.039539
## temperature.amplitude:temperature.min   1.243e-02  4.376e-03   2.841 0.004494
## total.cloud.amplitude:total.cloud.mean -1.277e-04  7.025e-05  -1.818 0.068999
## wind.speed.min.80:wind.dir.80           6.238e-04  2.686e-04   2.323 0.020203
##
## (Intercept)                            ***
```

```
## temperature.amplitude
## temperature.min
## pressure.amplitude                    *
## pressure.max                          ***
## total.cloud.amplitude                 .
## total.cloud.mean                      **
## med.cloud.amplitude                   ***
## med.cloud.min                         .
## wind.speed.min.80                     **
## wind.dir.80                           .
## wind.dir.900                          *
## wind.gust.amplitude                   *
## wind.gust.min                         *
## precipitation_boolTRUE                *
## snowfall                              *
## temperature.amplitude:temperature.min  **
## total.cloud.amplitude:total.cloud.mean .
## wind.speed.min.80:wind.dir.80         *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1296.85  on 935  degrees of freedom
## Residual deviance:  982.75  on 917  degrees of freedom
## AIC: 1020.7
##
## Number of Fisher Scoring iterations: 4
```

Le modèle identifié présente les caractéristiques suivantes

- 18 covariables
- 19 coefficients
- AIC = 1020.7

L'ajustement met en avant les covariables suivantes

- la **pression atmosphérique**
- la **température**
- la **nébulosité totale et moyenne** dont l'amplitude et le minima
- la **vitesse et direction du vent à 80 et 900m**
- les **rafales de vent**
- Les **précipitations** sous la forme booléenne
- l'**enneigement**
- des interactions

4.3.5 méthode progressive descendante

On procède à une sélection de covariables par **méthode pas-à-pas progressive descendante** (step both) à partir du modèle complet.

```
s2.res.step_both_from_full <- step (s2.res.glm.0,
                                    direction="both")
```

Le modèle identifié par la **méthode pas-à-pas progressive descendante** est le suivant

```
## pluie.demain ~ temperature.amplitude + temperature.min + pressure.amplitude +
##     pressure.max + total.cloud.amplitude + total.cloud.mean +
##     med.cloud.amplitude + med.cloud.min + wind.speed.min.80 +
##     wind.dir.80 + wind.dir.900 + wind.gust.amplitude + wind.gust.min +
##     precipitation_bool + snowfall + temperature.amplitude:temperature.min +
##     total.cloud.amplitude:total.cloud.mean + wind.speed.min.80:wind.dir.80
```

Son ajustement :

```
##
## Call:
## glm(formula = pluie.demain ~ temperature.amplitude + temperature.min +
##     pressure.amplitude + pressure.max + total.cloud.amplitude +
##     total.cloud.mean + med.cloud.amplitude + med.cloud.min +
##     wind.speed.min.80 + wind.dir.80 + wind.dir.900 + wind.gust.amplitude +
##     wind.gust.min + precipitation_bool + snowfall + temperature.amplitude:temperature.min +
##     total.cloud.amplitude:total.cloud.mean + wind.speed.min.80:wind.dir.80,
##     family = binomial, data = dat.meteo.train[scp.train, ], na.action = na.exclude)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2652  -0.8478   0.3199   0.8290   2.8122
##
## Coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                              6.757e+01  1.298e+01   5.205 1.94e-07
## temperature.amplitude                    3.438e-03  4.897e-02   0.070 0.944031
## temperature.min                         -6.540e-02  4.061e-02  -1.610 0.107338
## pressure.amplitude                       6.002e-02  2.802e-02   2.142 0.032203
## pressure.max                            -6.908e-02  1.258e-02  -5.490 4.02e-08
## total.cloud.amplitude                    7.666e-03  4.466e-03   1.717 0.086064
## total.cloud.mean                         1.835e-02  6.630e-03   2.768 0.005637
## med.cloud.amplitude                      9.937e-03  2.652e-03   3.748 0.000179
## med.cloud.min                            1.606e-02  9.602e-03   1.673 0.094409
## wind.speed.min.80                       -1.873e-01  6.685e-02  -2.802 0.005075
## wind.dir.80                             -4.113e-03  2.104e-03  -1.955 0.050599
## wind.dir.900                             2.837e-03  1.445e-03   1.963 0.049663
## wind.gust.amplitude                      1.720e-02  7.638e-03   2.253 0.024289
## wind.gust.min                            5.140e-02  2.232e-02   2.303 0.021297
## precipitation_boolTRUE                   4.776e-01  2.135e-01   2.237 0.025286
## snowfall                                -4.514e-01  2.193e-01  -2.059 0.039539
## temperature.amplitude:temperature.min    1.243e-02  4.376e-03   2.841 0.004494
## total.cloud.amplitude:total.cloud.mean -1.277e-04  7.025e-05  -1.818 0.068999
## wind.speed.min.80:wind.dir.80            6.238e-04  2.686e-04   2.323 0.020203
##
## (Intercept)                             ***
## temperature.amplitude
## temperature.min
## pressure.amplitude                      *
## pressure.max                            ***
## total.cloud.amplitude                   .
## total.cloud.mean                        **
## med.cloud.amplitude                     ***
## med.cloud.min                           .
## wind.speed.min.80                       **
## wind.dir.80                             .
## wind.dir.900                            *
## wind.gust.amplitude                     *
## wind.gust.min                           *
## precipitation_boolTRUE                  *
## snowfall                                *
## temperature.amplitude:temperature.min   **
## total.cloud.amplitude:total.cloud.mean  .
## wind.speed.min.80:wind.dir.80           *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1296.85  on 935  degrees of freedom
## Residual deviance:  982.75  on 917  degrees of freedom
## AIC: 1020.7
##
## Number of Fisher Scoring iterations: 4
```

Le modèle identifié par la méthode progressive descendante est identique à celui identifié par la méthode

descendante.

Il présente les caractéristiques suivantes

- 18 covariables
- 19 coefficients
- AIC = 1020.7

L'ajustement met en avant les covariables suivantes

- la **pression atmosphérique**
- la **température**
- la **nébulosité totale et moyenne** dont l'amplitude et le minima
- la **vitesse et direction du vent à 80 et 900m**
- les **rafales de vent**
- Les **précipitations** sous la forme booléenne
- l'**enneigement**
- des interactions

5 Validation des modèles

5.1 Mesure de la capacité prédictive des modèles

Pour chaque modèle candidat ajusté précédemment,

- on prédit le jeu de validation
- on détermine le **seuil optimal** qui maximise la **précision** savoir

$$precision = \frac{True\ Positive + True\ Negative}{n_{validation}}$$

- on mesure aussi
  - l'**AUC** (Area Under the Curve) grâce à la fonction ***ROCR::performance***
  - la pertinence du modèle par rapport à un modèle nul M_0 et un modèle saturé M_{sat} sur la base des **déviances** fournies par l'ajustement du modèle par glm.
    La pertinence est donnée par la p-valeur de tests de Chi2 de la déviance résiduelle D_k (n-(k+1) degrés de liberté) et la déviance $D_0 - D_k$ (k+1 degrés de liberté) où D_0 est la déviance nulle entre M_0 et M_{sat}.
  - la **perte** à savoir la moyenne des distances entre la probabilité prédite $p_i = P(Y_i = 1)$ et la vraie valeur de Y_i (connu pour le jeu de validation)

$$perte = \frac{1}{n_{val}} \sum_{i=1}^{n_{val}} |p_i - Y_i|$$

5.2 Résultats et choix du modèle

Table 2: Résumé des ajustements

	nb.covariables	nb.coefficients	aic
stratégie 1			
s1.res.glm.0	41	52	1055.489
s1.res.step_forward	8	19	1019.951
s1.res.step_both_from_constant	8	19	1019.951
s1.res.step_backward	18	29	1016.680
s1.res.step_both_from_full	18	29	1016.680
stratégie 2			
s2.res.glm.0	30	57	1064.280
s2.res.step_forward	12	23	1022.121
s2.res.step_both_from_constant	12	23	1022.121
s2.res.step_backward	15	19	1020.746
s2.res.step_both_from_full	15	19	1020.746

Table 3: Résumé des déviances

	deviance.test.mk.versus.msat	deviance.test.m0.versus.mk
stratégie 1		
s1.res.glm.0	0.0568581	0
s1.res.step_forward	0.0671003	0
s1.res.step_both_from_constant	0.0671003	0
s1.res.step_backward	0.1137773	0
s1.res.step_both_from_full	0.1137773	0
stratégie 2		
s2.res.glm.0	0.0472679	0
s2.res.step_forward	0.0721544	0
s2.res.step_both_from_constant	0.0721544	0
s2.res.step_backward	0.0648282	0
s2.res.step_both_from_full	0.0648282	0

Table 4: Mesures prédictives

	seuil.optimal	precision	auc	erreur
stratégie 1				
s1.res.glm.0	0.58	0.7254098	0.7857527	0.3631164
s1.res.step_forward	0.54	0.7336066	0.7700941	0.3761790
s1.res.step_both_from_constant	0.54	0.7336066	0.7700941	0.3761790
s1.res.step_backward	0.57	0.7295082	0.7877688	0.3619645
s1.res.step_both_from_full	0.57	0.7295082	0.7877688	0.3619645
stratégie 2				
s2.res.glm.0	0.54	0.7090164	0.7864247	0.3631938
s2.res.step_forward	0.53	0.7336066	0.7746640	0.3718330
s2.res.step_both_from_constant	0.53	0.7336066	0.7746640	0.3718330
s2.res.step_backward	0.44	0.7131148	0.7793683	0.3701820
s2.res.step_both_from_full	0.44	0.7131148	0.7793683	0.3701820

Synthèse

- D'après l'analyse de la déviance, tous les modèles proposés sont plus pertinents qu'un modèle saturé (p-valeur $> 5\%$) et plus pertinents qu'un modèle nul M0 (p-valeur $< 5\%$; les valeurs étant très petites, elles ont été arrondies à 0 par l'affichage).
- Tous proposent des niveaux d'**erreur** similaires de l'ordre de 0.36-0.37, ainsi que des précisions équivalentes de l'ordre de 71 à 73%

Selon le critère AIC,

- le meilleur modèle de la stratégie 1 est issu de la méthode pas-à-pas progressive descendante
- le meilleur modèle de la stratégie 2 est aussi issu de la méthode pas-à-pas progressive descendante

Dans l'optique de la prédiction du jeu de données **meteo.test**, on optera cependant pour le **modèle de la stratégie 2, issu de la méthode pas-à-pas ascendante (s2.res.step_forward)**:

- il propose un modèle de taille raisonnable en termes de covariables et de coefficients (12 covariables et 23 coefficients)
- il propose un AIC proche du meilleur de la stratégie (1022 versus 1020)
- sa précision est la meilleure obtenue (73,3%)
- son erreur est dans la norme des modèles proposé (0.371)

C'est ce modèle qui sera testé avec le jeu de données **meteo.test**.

6 Prédiction du jeu de test

Le modèle dédié que l'on considère le plus apte à prédire le jeu de données **meteo.test** a été identifié.

Ce modèle ayant été entraîné sur 80% du jeu de données **meteo.train** et validé sur 20% du même jeu de données, nous allons procéder de la façon suivante

- dans un 1er temps, le modèle va être réajusté sur le jeu de données **meteo.train** complet
- dans un 2ème temps, ce modèle réajusté va être utilisé pour prédire les valeurs $p_i = P(Y_i = 1)$ pour chaque entrée du jeu de données **meteo.test** (dans sa version transformée afin d'aligner les variables à celles utilisées au moment de l'ajustement)
- enfin, on utilisera le **seuil optimal** associé au modèle pour décider les valeurs Y_i : TRUE ou FALSE

Le vecteur obtenu est concaténé aux données **meteo.test** initiales, dans une colonne **pluie.demain**, puis sauvegardé dans un fichier **meteo.test.prediction.csv** en vu de l'évaluation.

6.1 Entraînement sur le jeu de données complet

Le modèle est entraîné sur l'ensemble des données du jeu **meteo.train**.

```
final.model.name <- "s2.res.step_forward"
final.model.formula <- s2.res.step_forward$formula
```

```
final.model.res.glm <- glm(final.model.formula,
                           data=dat.meteo.train,
                           family="binomial")
```

```
##
## Call:
## glm(formula = final.model.formula, family = "binomial", data = dat.meteo.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3968  -0.8389   0.3471   0.8223   2.6800
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)             69.772782  11.699242   5.964 2.46e-09 ***
## med.cloud.amplitude      0.009210   0.002179   4.226 2.38e-05 ***
## pressure.max            -0.072030   0.011387  -6.326 2.52e-10 ***
## precipitation_boolTRUE   0.678218   0.188419   3.600 0.000319 ***
## month2                  -0.406755   0.364728  -1.115 0.264752
## month3                  -1.010771   0.363898  -2.778 0.005476 **
## month4                  -1.042868   0.389587  -2.677 0.007432 **
## month5                  -0.771097   0.408984  -1.885 0.059377 .
## month6                  -0.345694   0.453528  -0.762 0.445922
## month7                  -0.526886   0.476411  -1.106 0.268749
## month8                  -0.994875   0.467669  -2.127 0.033395 *
## month9                  -1.376563   0.428277  -3.214 0.001308 **
## month10                 -1.048466   0.384151  -2.729 0.006347 **
## month11                 -0.868118   0.360427  -2.409 0.016015 *
## month12                  0.150719   0.367532   0.410 0.681744
## med.cloud.min            0.016103   0.008121   1.983 0.047383 *
## pressure.amplitude       0.075154   0.024176   3.109 0.001880 **
## snowfall                -0.447811   0.213696  -2.096 0.036122 *
## temperature.amplitude    0.161459   0.033910   4.761 1.92e-06 ***
## wind.dir.900             0.002753   0.001121   2.456 0.014054 *
## total.cloud.mean         0.012254   0.003875   3.162 0.001566 **
```

```
## temperature.min         0.047618   0.022231   2.142 0.032198 *
## wind.speed.min.900      0.006037   0.006900   0.875 0.381636
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1635.4  on 1179  degrees of freedom
## Residual deviance: 1247.7  on 1157  degrees of freedom
## AIC: 1293.7
##
## Number of Fisher Scoring iterations: 4
```

6.2 Prédiction et décision

Pour chaque individu du jeu de données **meteo.test**, on prédit $p_i = P(Y_i = TRUE)$

```
final.model.res.predict <- predict(final.model.res.glm,
                                   newdata=dat.meteo.test,
                                   type="response")
```

On procéde à la décision selon le **seuil optimal** identifié pendant la phase d'entraînement du modèle.

```
final.model.threshold <- RESULTS[final.model.name, RESULTS_current.model.seuil_opt]
final.model.prediction <- (final.model.res.predict >= final.model.threshold)
```

Les valeurs de *pluie.demain* prédites sont concaténées aux données initiales puis sauvegardées.

```
dat.meteo.test.prediction <- cbind(dat.meteo.test.raw, pluie.demain=final.model.prediction)
write.csv(dat.meteo.test.prediction, "meteo.test.prediction.csv",
          quote=FALSE,
          row.names = FALSE)
```

C'est ce fichier qui servira à l'évaluation des données prédites.